First International Workshop on
Intelligent Visual Interfaces for Text Analysis

Hong Kong, China
Feb 7, 2010

Proceedings

Edited by
Shixia Liu
Michelle X. Zhou
Giuseppe Carenini
Huamin Qu

ISBN: 978-1-60558-996-1

The Association for Computing Machinery
2 Penn Plaza, Suite 701
New York New York 10121-0701

Table of Contents

Papers

Workshop Committee

Shixia Liu
IBM China Research Lab, China

Michelle X. Zhou
IBM China Research Lab, China

Giuseppe Carenini
University of British Columbia, Canada

Huamin Qu
Hong Kong University of Science and Technology, Hong Kong

Program Committee

Ed Chi
Palo Alto Research Center, USA

Christopher Collins
University of Toronto, Canada

Jeffrey Heer
Stanford University , USA

Tomoharu Iwata
NTT Communication Science Laboratories, Japan

Chin-Yew Lin
MSRA, China

Shimei Pan
IBM Watson Research Center, USA

John T. Stasko
Georgia Institute of Technology, USA

Pak Chung Wong
Battelle Pacific Northwest Division, USA

HARVEST: An Intelligent Visual Analytic Tool for the Masses

David Gotz, Zhen When, Jie Lu, Peter Kissa
IBM T.J. Watson Research Center
19 Skyline Dr, Hawthorne, NY, USA

Nan Cao, Wei Hong Qian, Shi Xia Liu, Michelle X. Zhou
IBM China Research Lab
8 Dongbeiwang West Rd, Beijing, China

ABSTRACT
We present an intelligent visual analytic system called HARVEST. It combines three key technologies to support a complex, exploratory visual analytic process for non-experts: (1) a set of smart visual analytic widgets, (2) a visualization recommendation engine, and (3) an insight provenance mechanism. Study results show that HARVEST helped users analyze a corpus of text documents from a corporate wiki.

Author Keywords
Visual Analytics, Smart Graphics, Visualization

ACM Classification Keywords
Algorithms, Human Factors

INTRODUCTION AND RELATED WORK
In recent years, a large number of visualization systems have been developed to help users view, explore, and analyze information. The capabilities supported by these visualization systems vary broadly, ranging from supporting casual visual collaborations (e.g., ManyEyes [11] and Swivel [1]) to commercial-grade visual analytics (e.g., Spotfire [3] and Tableau [2]).

At the same time, businesses have been creating and storing more data than ever before. Recognizing that valuable insights are buried within these mountains of information, companies have begun to push the use of visualization to all their employees to aid their business decision-making processes. However, most of today's visualization tools still target two niche audiences: (1) dedicated information analysts and (2) dashboard consumers.

Tools for information analysts cater to users who have acquired a high degree of visualization and computer skills and often use sophisticated visualization software. However, they are typically too complex for average business users. In contrast, dashboard consumers are typically casual users of visualization. By design, dashboard systems require far less skill and are accessible to a much wider range of users. However, they lack several key capabilities, such as continuous exploration of large data sets, which are often required to support real-world business tasks.

However, there is a third and perhaps largest class of users for whom existing tools are of limited value: everyday business workers. These users typically have extensive domain knowledge but are not visualization or computer experts. Yet as part of their daily responsibilities, they perform situational analysis tasks over massive amounts of data for which visualization can be of great benefit.

For example, in our own company, employees often examine a large wiki site containing data about numerous projects underway within our organization. While the wiki effectively provides information on individual projects, it is very difficult for users to examine project patterns or trends. Neither can most existing visualization tools make this sort of task any easier for an average person.

To help this user population, we are building HARVEST, an intelligent visual analytic system for everyday business users. HARVEST combines three key technologies to support an exploratory visual analytic process without requiring users to be visualization or computer experts:

- **Smart visual analytic widgets.** A set of visualization widgets that can be easily reused across applications. They support semantics-based user interaction to help identify and capture user intention, and incrementally handle dynamic data sets retrieved during a continuous visual analytic task.
- **Dynamic visualization recommendation.** A context-driven approach that assists users in finding the proper visualizations for use in their context.
- **Semantics-based capture of insight provenance.** A semantics-based approach to modeling and capturing a user's logical analytic process. It supports automatic detection of user action patterns for better visualization recommendation, and enables flexible adaptation of a user's analytic process for reuse in new contexts.

Our work is related to previous systems that use automatic visualization (e.g., [9]) but focuses on situational analytics where role or template based approaches are less effective. Our work is also related to systems that capture user histories (e.g., [4, 8, 10]). However, HARVEST focuses on the extraction of a semantic representation of a user's insight provenance that is independent of application and across a range of visualization tools. It then analyzes that provenance to provide context-relevant visualization recommendations.

REFERENCE APPLICATION
Our work on HARVEST is motivated by the common information needs of employees within our own company. Our organization maintains a large wiki site describing all ongoing research projects. Each project page is a semi-structured

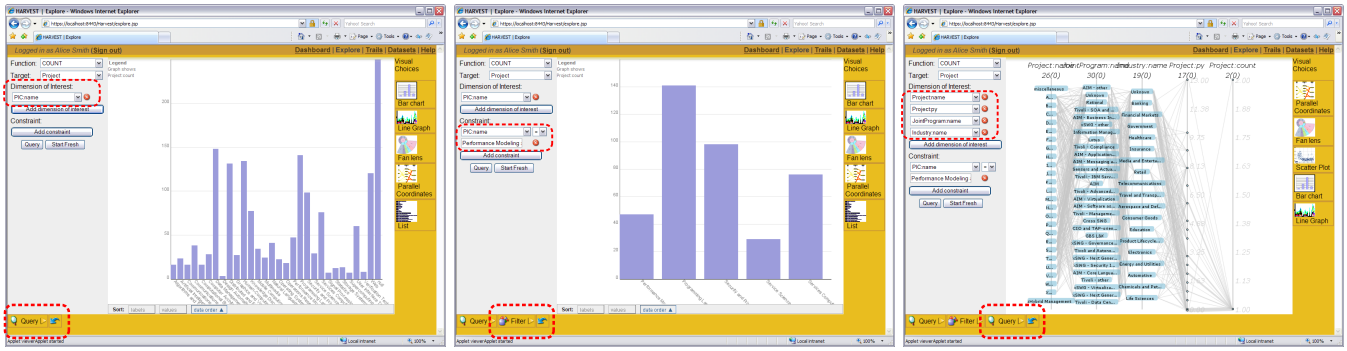


Figure 1. Screenshots illustrating the typical user workflow in our reference HARVEST application. In addition to the changing visualization canvas, user progress is reflected in both the query panel and history panel as we highlight with in red.

text document, containing a project description, the people involved, and several other important pieces of information. New projects are added to the wiki regularly, and updates are constantly contributed by project members and managers. While it is relatively easy to look up information about individual projects in the wiki, there is no easy way to obtain an overview of a collection of projects. Yet higher-level summaries of information may often be most valuable.

For example, consider a researcher named Alice who is putting together a new proposal for a computer vision research project. To scope her project properly, Alice must decide how many 'person-years' (PYs) could be realistically funded. To help answer this question, Alice would like to view the distribution of PYs in funded projects, especially in the area of computer vision. Similarly, Alice could better position her proposal if she could discover which funding programs were historically most likely to accept computer vision proposals. In addition, she would like to identify potential collaboration partners by examining related projects and their teams. The information required to answer each of Alice's questions is contained within the project wiki. However, there is no easy way for Alice to extract the needed insights. HARVEST is designed to help people like Alice by providing a set of intelligent visual analysis tools.

Before HARVEST can be used for this application, the wiki data was pre-processed by a text-analysis tool to extract key terms and concepts. The extracted data was then stored together with the documents' structured meta-data within a DB2 database. Then, Alice begins by logging in to HARVEST and initiating a new task. She starts by using the query GUI panel to build a query to summarize the number of projects by discipline. In response, a *Query* action is processed by the three core HARVEST components: (1) the *query manager* interprets the GUI input to formulate a SQL query and executes it, (2) the *visualization recommender* automatically composes a bar chart encoding the retrieved data, and (3) the *action tracker* incorporates the *Query* action into its representation of Alice's insight provenance. The visualization and the newly performed *Query* action are displayed in Figure 1(a).

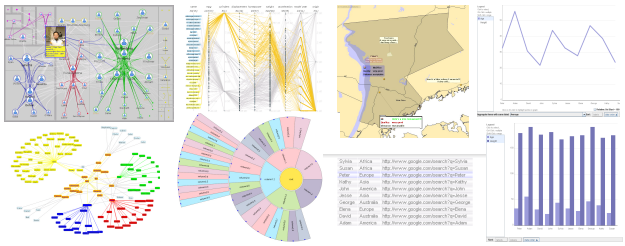


Figure 2. Samples of visual analytic widgets used in HARVEST.

Alice then selects a subset of five bars from the visualization that correspond to five disciplines in which she is interested. She issues a *Filter* action using the bar chart's context-sensitive menu. In response, HARVEST updates the visualization to reflect Alice's new data interests. Both the query and history panels also are updated to include the new data constraints and the Filter action, respectively (Figure 1b).

For all projects in the five selected disciplines, Alice now wants to examine the correlations among four variables: the discipline, funding partner, project PY, and related industry. To do so, Alice modifies the current query and submits it. In response to this new *Query* action, HARVEST creates a parallel coordinates visualization to encode the updated data (Figure 1c). After identifying an important trend, Alice switches to a list view of the documents and selects individual items to view the full text. As shown here, Alice's analysis goals and data interests evolve over the course of her task, making it impossible to know ahead of time which data sets Alice would like to analyze or the proper visualizations to use.

KEY HARVEST TECHNOLOGIES

HARVEST combines three key technologies: (1) smart visual analytic widgets (Figure 2), (2) dynamic visualization recommendation, and (3) semantics-based capture of insight provenance.

Smart Visual Analytic Widgets

HARVEST's smart visual analytic widgets support incremental visual updates to accommodate users' evolving data interests due to the exploratory nature of their tasks. Few existing visualization tools support incremental updates. In-

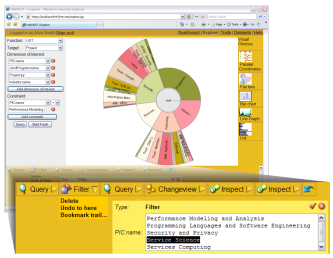


Figure 3. The history panel displays the unfolding analytic trail.

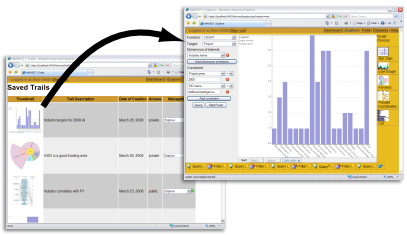


Figure 4. Users can restore saved trails to reuse past analyses.

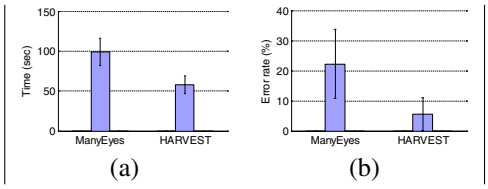


Figure 5. Mean and 95% confidence interval of (a) task completion time and (b) task error rate.

stead, a new visualization must be created if the underlying data changes. However, the abrupt change when creating a new visualization disrupts visual continuity and reduces a user's ability to comprehend information across successive displays [12]. To address this issue, a subset of our visual widgets is designed to support incremental visual updates. They include a visual context management module to dynamically decide how to best update the existing visualization to incorporate new data [12]. For example, When a user issues a follow-up query to retrieve additional data, our SmartMap widget dynamically derives a set of visual animation operators to incrementally update the existing visualization, such as *CameraSwitch*, *Add* (adding visual representations of the new data), and *Simplify* (visually simplifying visual representations of old data).

In addition, each widget supports a set of semantics-based user actions to capture the semantics of a user's insight provenance. One of HARVEST's key goals is to capture the semantics of insight provenance, which can be used to help share and re-purpose a user's visual analytic processes. Since a large part of a user's activity is interacting with visual widgets, ideally these widgets should recognize the semantics of user activities as they occur. This is in contrast to most existing visualization tools that support an event-based interaction model (e.g., clicks and drags) and know little about the semantics of a user activity. To achieve our goal, we implement visual analytic widgets to support a set of actions, which are semantics-based interaction primitives (see Section *Semantics-Based Capture of Insight Provenance*).

Dynamic Visualization Recommendation

As demonstrated in our reference scenario, it is impossible to determine ahead of time which visualization tools should be used. To assist average users in effectively using visualizations in their tasks, we develop a visualization recommendation engine. Given a user's request, our engine automatically recommends the top-N suitable visualizations to the user. This engine extends our previous effort in automated visualization generation [13], which was limited to handling small data sets in a relatively static environment. To support continuous user interaction with large data sets in real-world HARVEST applications, we have extended our work to consider user behavior.

During visual analysis, a user's behavior can often signals implicit analytic needs [7]. For example, assume that Alice is interacting with a FanLens that hierarchically encodes the number of projects by discipline and by sponsor (Figure 3). To compare the number of projects by sponsor in each of the disciplines, she iteratively clicks on each discipline (a slice) to expand it. To better help Alice work more effectively in the above situation, HARVEST provides behavior-based visualization recommendation [5]. First, as a user interacts with HARVEST, the action tracker component examines the user's action history in search of meaningful patterns (see the *Automatic Pattern Detection in User Actions* section). Once a pattern is detected, we use a rule-based approach to map the pattern to an implied visual task. For example, the pattern demonstrated by Alice is mapped to a visual comparison task. Based on the inferred visual task, our visualization recommendation engine would recommend a bar chart visualization to Alice that more effectively encodes the desired information for direct comparison.

Semantics-Based Capture of Insight Provenance

Visual analytic tasks are often complex and time consuming. To make the process easier for average business users, HARVEST's action tracker component maintains a semantics-based model of a user's visual analytic activity. This model is then used to enable more effective visualization recommendation, and to allow flexible adaptation of a user's analytic process to new situations. We refer to the model of user activity as insight provenance because it contains the history and rationale of how insights are derived during a user's visual analytic process.

Automatic Identification of User Analytic Trails

Our empirical studies [7] demonstrate that distinct logical sequences of user actions leading to an insight, which we call analytic trails, can be observed in a user's analysis process. Using our reference application scenario, Alice has performed three actions, *Query* $\Rightarrow$ *Filter* $\Rightarrow$ *Query* to reach the state shown in Figure 2(c). As this example illustrates, trails define a user's exploration path and its semantics (e.g., captured by the action types and parameters).

HARVEST actively analyzes the linear sequence of user actions in the order they are performed by an analyst. Based on the type of action being performed, HARVEST builds a graph-based representation of interconnected trails to represent the user's visual exploration behavior. When users save their work via *Bookmark*, HARVEST preserves both the state of the visualization as well as the automatically recorded analytic trail. When a bookmark is later restored, the trail is restored as well. This allows a user to review the exploration

context in which an insight was discovered. This feature is especially useful during collaborative tasks, allowing users to see not only *what* a coworker has found, but also *how* they found it.

Automatic Pattern Detection in User Actions

In addition to identifying analytic trails, the action tracker performs pattern detection over a user's recently performed actions in search of meaningful activity patterns. We define a pattern as an iterative user behavior that implies a specific analytic goal. Studies show that patterns occur frequently in typical visual analytic behavior and correlate with real or perceived limitations in a tool [7]. Detected patterns are passed as input to the visual recommender to enable user behavior-driven recommendations. HARVEST uses a rule-based approach to pattern detection. Each time a user performs a new action, the user's analytic trail is compared against a library of pattern rules. The library includes one or more rules for each pattern recognized by HARVEST. The system currently detects four user action patterns: *Scan*, *Flip*, *DrillDown*, and *Swap* [5].

Flexible Adaptation of Analytic Trails

One of the main benefits of HARVEST's automatic capture of analytic trails is that it allows users to adapt their previous analysis processes to new tasks. As a user interacts with the system, HARVEST externalizes a simplified version of the user's exploration path in the history panel (Figure 3). Users can interact with the history panel to directly manipulate their trail. Supported manipulations include: *Undo*, *Revisit*, *Delete*, and *Modify*. This allows users to quickly adapt their previously performed trails to new contexts. User's can modify specific parameters of an action (e.g., change a *Filter* from $Year = 2008$ to $Year = 2006$). Whenever the user's analytic trail is altered, HARVEST automatically requeries for new data and composes an updated visualization. These capabilities are especially powerful when combined with bookmarks. Rather than always starting from scratch, a user can make use of previously saved trails from similar tasks. After restoring a saved trail, a user can back up to any action in the restored trail to use as a starting point for his/her new analysis. Alternatively, s/he can re-use the entire trail and simply modify individual action parameters to meet the new needs.

EVALUATION

We applied HARVEST to the reference application described earlier and conducted a user study with eight participants, using a modified version of ManyEyes [11] as a baseline. Figure 5 shows that HARVEST performed significantly better by our two objective metrics: task completion time and error rate. Users of HARVEST were able to complete their tasks significantly faster ($p < 0.0001$), with about 40% time reduction at each step of the task on average (Figure 5a). Note that we ignored the time spent retrieving, formatting, and uploading data as required by ManyEyes. Thus, the time accounted for was spent by a user to select a visualization, interact with the visualization to analyze the information, and switch to a new visualization if needed. We attribute this significant reduction in time mainly to HARVEST's visualization recommendation, which quickly led users to proper visualizations for their tasks.

Our results also indicated a significant difference in task error rate between the two systems ($p < 0.01$) (Figure 5b). When we checked user results with facts from the original content, we found that there was a 75% reduction in error rate on average when a task was performed using HARVEST (5.6%) vs. using ManyEyes (22%). We attribute the sharp drop in error rate to HARVEST's ability to let users easily explore data from different angles. Users commented that HARVEST made it "*easy to switch*" to alternative visualizations and different data sets. Moreover, HARVEST's automated analytic trail management facility made operations like "go back" or "undo" trivial. In essence, it was the seamless integration of HARVEST's key technologies that led to more accurate results. As one user commented, "*[there was] coordination among [the] query GUI, analytic trail, and visualization [in HARVEST]...*", where you could "*modify/specify queries from any of the three.*" Finally, from users' subjective feedback, the participants also overwhelmingly favored HARVEST (mean rating of 4 out of 5) over ManyEyes (mean rating of 2.6) for the tasks that they performed.

CONCLUSION

In this paper, we have presented HARVEST, an intelligent visual analytic system designed to empower everyday business users to derive insight from large amounts of data. We reviewed the key technologies behind the HARVEST system and presented results from a user study. Our study shows that HARVEST technologies were preferred by users, and that they helped users perform significantly better by two objective metrics: task completion time and error rate.

REFERENCES

1. Swivel. http://www.swivel.com/.
2. Tableau software. http://www.tableausoftware.com/.
3. Tibco spotfire. http://spotfire.tibco.com/.
4. L. Bavoil and et al. Vistrails: Enabling interactive multiple-view visualizations. In *IEEE Vis*, 2005.
5. D. Gotz and Z. Wen. User behavior driven visualization recommendation. In *IUI 2009*.
6. D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. In *IEEE VAST*, 2008.
7. D. Gotz and M. X. Zhou. An empirical study of user interaction behavior during visual analysis. Technical Report RC24525, IBM Research, 2008.
8. M. Kreuseler, T. Nocke, and H. Schumann. A history mechanism for visual data mining. In *IEEE InfoVis*, 2004.
9. J. D. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. In *Proc. of IEEE InfoVis*, 2007.
10. Y. B. Shrinivasan and J. J. van Wijk. Supporting the analytical reasoning process in information visualization. In *CHI*, 2008.
11. F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Many eyes: A site for visualization at internet scale. In *InfoVis*, 2007.
12. Z. Wen, M. X. Zhou, and V. Aggarwal. An optimization-based approach to dynamic visual context management. In *InfoVis*, 2005.
13. M. X. Zhou and M. Chen. Automated generation of graphical sketches by example. In *Proceedings of IJCAI*, pages 65–74, 2003.

A Dynamic Visual Interface for News Stream Analysis

Weiwei Cui, Hong Zhou, Huamin Qu
Hong Kong University of Science and Technology
{weiwei, hongzhou, huamin}@cse.ust.hk

Wenbin Zhang, Steven Skiena
State University of New York at Stony Brook
{wbzhang, skiena}@cs.sunysb.edu

ABSTRACT
In this paper, we introduce a new visualization primitive called TextWheel, and present a visual analytics system for news streams which can bring multiple attributes of the news articles and the macro/micro relations between news streams and keywords into one coherent analytical context and meanwhile convey the dynamic natures of news streams. We use our system to analyze several large-scale news corpora related to some major companies and the results demonstrate the high potential of our method.

Author Keywords
Dynamic Visualization, Large-Scale Document Visualization, Time Series Data

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Optional sub-category*

INTRODUCTION
Keyword-based searching and clustering have been widely used for news analysis (e.g.,[1]), because, for topics such as major companies or events, the enormous volume of news articles makes reading them one by one basically impossible. On the other hand, there may exist complex macro/micro relations between keywords and articles. At the micro level, keywords (e.g., Bill Gates and Microsoft) have various relations. At the macro level, articles may be also related (e.g., dealing with the same topic). Meanwhile, each article contains multiple keywords and each keyword may appear in many articles. These complex macro/micro relations may be very useful for text analysis. However, it is very difficult to visually explore them for a news stream. First, news articles may have many attributes, such as author, time, length, and sentiment, which should be taken into account. It is a classic hard problem to develop visual encoding schemes for multivariate and time-varying attributes. Second, the large size of news streams may also pose special challenges for the scalability of visual encoding schemes.

In this paper we develop a visual analytics system for news streams and try to address the above-mentioned issues. We focus on the multiple attributes of news articles and the dynamic relations between articles and keywords.

To deal with the multiple attributes of keywords, we first perform natural language processing on the news articles. The keywords in them are identified. Meanwhile their various attributes (e.g., sentiment and frequency) are also obtained. After that, attributes are collected and summarized as a line chart called significance trend chart to provide an overview of the attribute evolution over time. At the same time, a concise glyph is used to encode individual article with multiple attributes using different visual channels of the glyph to provide an overview of the article.

To deal with the complex macro/micro relations among keywords and articles, we introduce a novel visual primitive called TextWheel which consists of one or multiple keyword wheels, a document transportation belt, and chains to connect the belt and wheels (see Fig. 1). By observing the TextWheel and its content changes, interesting patterns can be detected. Our method is intuitive, and can be easily extended to visualize other text files (e.g., blogs, emails,internal memos, and fictions), or even other data formats such as video clips.

DATA PROCESSING

Entity Recognition and Sentiment Analysis
Named entity recognition is a well-studied problem with an extensive literature (e.g. [6]). We primarily employ rule-based techniques which are not vastly different than those in the literature, although they require substantial engineering to achieve good performance. Interested users can refer to [3]. In this recognition process, we label different entities with different categories, such as company, person, and country, and collect their frequency information.

On the other hand, *Sentiment analysis* of texts is a large and growing field, surveyed in [4]. News articles often express opinion of news entities (people, places, etc.), which can be very useful for text analysis. Inspired by the method introduced in [2], we have developed a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus. By tracking reference frequencies to adjectives with positive and negative connotations, we evaluate $entity\_polarity_i$ using sentiment counts data for as

$$entity\_polarity_i = \frac{positive\_sentiment\_references_i}{total\_sentiment\_references_i}$$

Significance Trend Chart
In this sub-section, we introduce significance trend chart, which is inspired by entropy and information theory, to analyze and visually summarize the change of sentiment and word frequency information in the whole document stream.

According to entropy and information theory, if an object contains more exclusive information, it is more significant. Following this idea, we define that a document is more significant if it has more exclusive sentiment information, compared with its neighboring (preceding or succeeding) documents in the document stream. Therefore, we estimate the the significance value $S(X_t)$ of document X_t at time t as:

$$S(X_t) = \frac{1}{2}(2H(X_t) - H(X_t;X_{t-1}) - H(X_t;X_{t+1}))$$

where $H(X_t)$ is the entropy value of document X_t, which measures the amount of information in X_t, $H(X_t;Y)$ is the mutual information value of X_t (given another document Y), which measures the amount of information shared between X_t and Y, X_{t-1} and X_{t+1} are the document X_t's preceding and succeeding documents, respectively.

To calculate $H(X)$, a histogram is built upon the average sentiment values towards every keyword in the document.Based on the histogram, $H(X)$ of the document can be computed by using the normalized count of every histogram bin, i.e.,

$$H(X) = -\sum_{i=1}^{N} \frac{\text{cnt}_i}{\text{cnt}_X} \log \frac{\text{cnt}_i}{\text{cnt}_X}$$

where N is the bin number, and cnt_i and cnt_X are the word counts in the ith bin and the whole document$X$, respectively.

Similarly, given another document Y, $H(X;Y)$ of document X can be calculated based on their joint histogram:

$$H(X;Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\text{cnt}_{ij}}{\text{cnt}_{XY}} \log \frac{\text{cnt}_{ij} \cdot \text{cnt}_X \cdot \text{cnt}_Y}{\text{cnt}_i \cdot \text{cnt}_j \cdot \text{cnt}_{XY}}$$

where N is the bin number in X histogram, M is the bin number in Y histogram, cnt_{ij} is the number of words which fall in both ith bin in X histogram and jth bin in Y histogram, cnt_{XY} is the number of words shared by document X and Y.

TEXTWHEEL

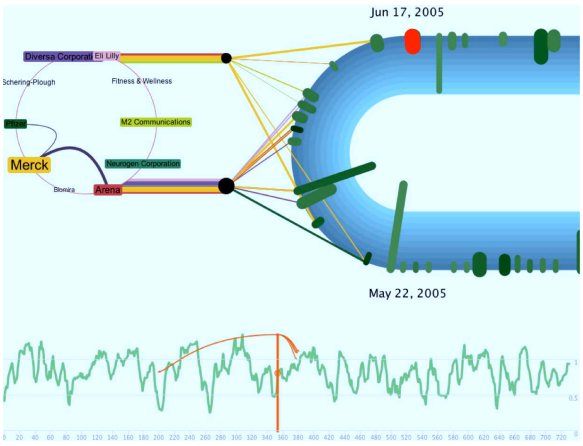


Figure 1. TextWheel interface: visualization of news streams to reveal multiple attributes of news articles and the macro/micro relations.

Fig. 1 shows the interface of our TextWheel visualization system with all its main visual components: significance trend chart, one document transportation belt, one or multiple keyword wheels. These three components provide users with a coherent three levels of details in the news stream.

At the bottom, the significance trend chart depicts the sentiment changes, which provides users the highest level of view. The x-axis encodes the time and the y-axis encodes the significance value of the documents.

On the upper right, the U-shape document transportation belt shows users a small portion of documents in the whole news stream. Each glyph on it represents one or multiple documents. During the process of exploration, the document glyphs may be transported along the belt. Users can closely analyze a small set of documents falling into a focus region on the belt (the semicircular part of the transportation belt) while the other document glyphs serve as the context. When a new document glyph enters the focus region, it is highlighted by using red color (see Fig. 1). Compared with other focus+context techniques, the U-shape belt can better convey the "endless" feeling about news streams. By using the U-shape, users can also easily compare the documents before and after the focus region side-by-side.

To synchronize the transportation belt and the significance trend chart, we put a sliding bar on the chart to show the location of the highlighted document glyph in the whole document stream. We also encode the macro-relations between documents on the chart. By drawing arcs from the sliding bar, all the documents which are most related to the highlighted document are pointed out in the whole document stream. Therefore, if users find those documents are interesting, they can conveniently drag the sliding bar to those locations for further exploration.

On the upper left, one or two wheel show different keywords users are interested in. To encode the micro-relations between keywords, we position keywords uniformly into a circular frame and then use lines to indicate the relations between keywords. The keyword wheels can interact with all the documents in the focus area by connecting them with some chains. These chains are similar to the edges in bigraphs which are widely used to encode the relations similar to the one between keywords and documents.

Entity Encoding with Glyphs

Keyword Glyph As keywords have been widely used in web pages, some visual encoding schemes have been well established. We adopt the tag cloud scheme in the system. The text sizes encode word frequencies in all the documents in the focus area. And the colors are uniquely assigned for each word in the same wheel for distinction purpose.

Document Glyph We adopt a simple rectangular shape glyph in our current system, though other more complicated glyphs can also be used. The width of the rectangle encodes the average number of keywords while the height indicates the article length. The color of the glyph encodes the average

sentiment expressed in an article. Based on the encoding scheme, our system can automatically change the appearance of the glyphs according to the associated data attributes.

Document Transportation Belts

Speed The speed of the transportation belt can be controlled. There are two ways to set the speed of the belt: automatically computing and manually setting. We can either set a uniform speed for the transportation belt or compute the speed based on the significance trend chart. We can also directly drag the sliding bar on the significance trend chart to fast forward or backward the belts to the time we are interested in. If users find something interesting, they can stop the belt or lower the speed to allow more time for inspection.

Order of documents The documents can come into the transportation belt with different orders. By default, they are arranged by time. Other orders are also possible. For example, the documents from the same sources can come together.

Keyword Wheels

Keyword selection We select the most relevant keywords from the documents collected in a larger time scale. For example, if the window region will show the documents in one month, the keywords will be selected based on the statistics from a one-year period. The keyword glyphs are put in the keyword wheels. The relation between two keywords is encoded by simply connecting them with a line.

Keyword position Keywords will be uniformly positioned in the circular frame of the keyword wheel. Their positions are computed based on their inter-relations. If two words are more correlated to one another, they are more likely to be placed together on the wheel.

Keyword update As the documents in the focus window change gradually, some keywords may become more or less frequent, or even disappear from the documents in the focus window. If a keyword becomes more frequent, the glyph size becomes bigger, and vise verse. When a keyword disappears from the documents in the focus window, it is still kept on the wheel. Its background color, however, will disappear, so that it will not cause much distraction to users.

Dynamic System

We further connect the keyword wheel and the document transportation belt with chains to encode relations between keywords and documents. Once a document enters the focus window, a chain will connect the document with each related keyword in the keyword wheel. The width and color of the chain can encode various attributes of the relation between the document and the keyword. For example, we can use the width to encode the strength of the sentiment and use the color to indicate what word this sentiment is about. For each wheel, we put two hubs between the keyword wheel and the transportation belt. Every chain will go through one of them first. In our system, all the chains indicating positive sentiments go to the lower hub, while all the chains indicating negative sentiments go to the upper hub. At each hub, all the chains with same color are bundled together and connected to the keyword wheel to drive the wheel and cause it to rotate. We assume both bundled chains have attractive forces on the wheel, and the force values are proportional to the bundle widths. In other words, if the sum of negative sentiments towards to all the keywords is stronger than the sum of positive sentiments, the wheel will rotate clockwise in our system, and vise verse.

CASE STUDIES

We have applied our system to a financial news corpus. This news corpus contains 333,289 articles published between 2004 and 2006, and relating to six major topics: Microsoft, Sony, NYSE, Merck, China, and Verizon. Each topic contains thousands of news articles from various sources. We first ran our system for each topic and did some initial screening. Once we found interesting patterns, we configured the system and fine tuned the visual displays to bring out more details for analysis. In this section, we describe one case study regarding to the topic *Merck*. The experiment was conducted on a Macbook Pro with Intel Core 2 Duo 2.2GHz CPUs and 2GB Memory.

For the news streams related to Merck, we divided the keywords into two groups, i.e., company and drug, and hoped it could shed light on the drugs Merck marketed and the other companies Merck had relations with. First, we quickly dragged the sliding bar and scanned the whole news sequence. We then noticed that a drug called Vioxx appears frequently in the display and the sentiment toward it changes dramatically at 2004 from neutral to bad. Then we reran the system and paid special attentions to this drug. Fig. 2 shows some screen shots. According to Fig.2 (a), Vioxx has not appeared in the drug keyword wheel until around Aug. 23, 2004. Then on Aug. 30, Vioxx shows up and the sentiment toward it is quite negative (see Fig. 2 (b)). We followed the link from the Vioxx glyph, identified a slim document glyph representing articles on Aug. 26. This article from the AFX UK Focus wrote: “Analysis of a study on the safety of COX-2 inhibitors found that Vioxx doses above 25 milligrams per day tripled the risk of cardiovascular...”. Therefore, Merck, the maker of Vioxx, also has slightly stronger negative sentiment than positive sentiment. Both Merck and Vioxx keep stable until Oct.1, when the Vioxx becomes much more negative (see Fig. 2 (c)). We followed the links from the Vioxx glyph, and retrieved the corresponding articles. One article from the AFX International Focus mentioned that “Merck said Merck was withdrawing its Vioxx arthritis drug from shelves worldwide, resulting in a 50 cent-to-60 cent reduction in per-share earnings.” We then followed the arcs on the significance trend chart to reveal similar news articles from the same source. (see red lines in Fig. 2 (c)). This article also said something negative about Merck - “Merck slumped more than 5 percent and was the biggest percentage loser among Dow Jones Industrial Average components.”

In the keyword wheel, we found that there is a strong relation between Pfizer and Merck (see Fig. 2 (a) and 2 (b)). It turned out that these two companies are competitors. After Aug. 2004, Celebrex and Vioxx start to appear together frequently in the news streams (see Fig. 2 (b) and 2 (c)). We re-

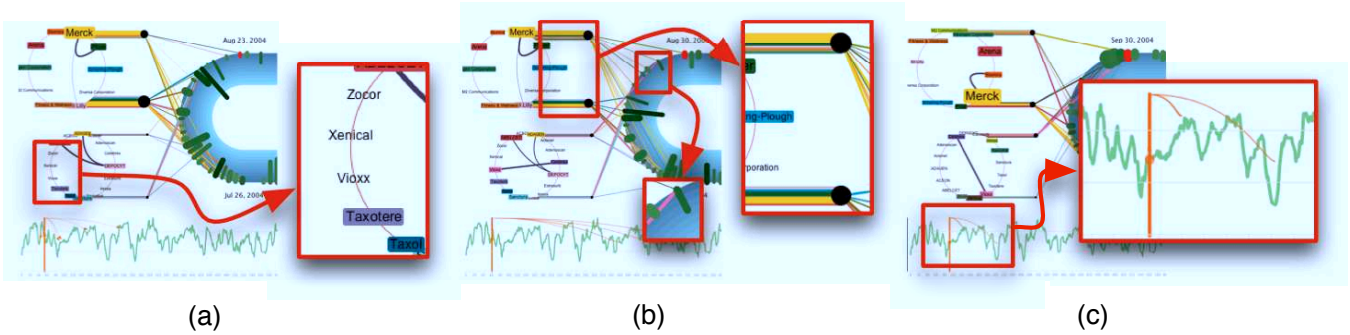


Figure 2. The story of Merck and its troublesome drug Vioxx.

trieved a related article and it wrote "Pfizer sank on increased speculation that its drug Celebrex may cause the same cardiovascular problems as Merck's recalled drug Vioxx." It became clear that Celebrex was a drug manufactured by Pfizer and was also in doubt at that time. This case study demonstrates that the sentiments expressed by the keyword glyphs are very useful in news analysis and with the macro/micro relation information provided by our system we can quickly identify the sources of those sentiments.

USER STUDY

In addition to the case study, we also conducted a user study and invited 12 college students to participate. After familiarizing them with our system, we asked them to finish two tasks and recorded their answers and response time. These two tasks are designed to test the effectiveness of keyword wheel and micro relation encoding scheme respectively. (The effectiveness of macro/micro relation encoding scheme is illustrated in the case study.) Both tasks are performed on a document corpus with 1,616 documents in one year.

In the first task, the users were asked to discern when a specific keyword on the wheel reaches its largest positive sentiment. In the second task, the users were asked to discern which two keywords have strongest relationship in the whole document stream.

For the first task, 10 users successfully found the correct time, while the rest found the time where it is also a sentiment peak with second largest positive sentiment. The average response time was 35 second. However, we also noticed that the standard derivation was as big as 21 seconds, which is probably because some of them were still not quite familiar with our system. The second task is a little harder than the first one, since users may need to track multiple keywords at the same time. This time, 8 users found the correct pair. The average response time was 54 second. However, The standard derivation was reduced to 18 seconds. We can see that the users were getting more and more familiar with our system. These two tasks demonstrates that with, a little training, most users can use our system to explore large news streams and correctly find pattern in the testing data.

In addition to the tasks, we also asked the users about their general feeling about our design. The most concern we have is the visual clutter and distraction in the interface. However, the responses to our system from users are quite enthusiastic. They feel the system is informative, intuitive, and visually appealing. We plan to deploy to a news website owned by our collaborators to reach more audiences and further improve our system based on their comments.

CONCLUSION

In this paper, we have presented a visual analytics system for large-scale news streams. Our system provides the multiple attributes of news articles and keywords, the dynamical macro/micro relations between news articles and keywords. To the best of our knowledge, it is the first time these useful information can be encoded and analyzed using one display. Meanwhile, we also identify some weaknesses of our system. For example, too many glyphs may overwhelm users and clutter the display. To deal with this problem, it is better to use our system together with some data mining techniques to first narrow down the document scope. In the future, we plan to further extend our system to encode more attributes of text documents. We believe the integration of our system with other data mining methods will make it more powerful. Encoding the uncertainty associated with the sentiment and co-occurrence computations is worth further study.

REFERENCES

1. J. A. Fitzpatrick, J. Reffell, and M. Aydelott. Breakingstory: visualizing change in online news. In *extended abstracts on Human factors in computing systems*, pages 900–901, 2003.

2. N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *the Intl. Conf. on Weblogs and Social Media*, 2007.

3. C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT Press, 2002.

4. B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers, 2008.

5. G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480, Philadelphia, Pennsylvania, USA, 2002.

Visualizing Common Sense Connections with Luminoso

Robert Speer
MIT Media Lab
Cambridge, MA, USA
rspeer@mit.edu

Catherine Havasi
MIT Media Lab
Cambridge, MA, USA
havasi@mit.edu

Nichole Treadway
MIT EECS
Cambridge, MA, USA
knt@mit.edu

Henry Lieberman
MIT Media Lab
Cambridge, MA, USA
lieber@media.mit.edu

ABSTRACT
We present Luminoso, a tool that helps researchers to visualize and understand a dimensionality-reduced semantic space based on textual information by exploring it interactively. It streamlines the process of creating such a space by taking input from a directory of text documents, and optionally including common-sense background information. This interface is useful for interactively discovering trends in a text corpus, such as free-text responses to a survey. We discuss a case study about restaurant reviews to show how Luminoso can be used for opinion mining.

Author Keywords
n-dimensional visualization, common sense, svd, natural language processing

ACM Classification Keywords
I.6.9 Simulation, Modeling, and Visualization: Visualization

INTRODUCTION

When people express their opinions in large quantities through surveys, forums, comments, and dialogue systems, these opinions contain a wealth of information that can be difficult to extract. In order to use this information, we need an interface that allows us to naturally interact with the textual data we've collected. The interface should create and intelligently present a space that displays the large-scale patterns and distinctions in a corpus of textual data. On top of this, the modes of interaction with the interface should be intuitive enough that a user can understand the information it is presenting without having specialized knowledge about machine learning.

Our goal is to provide a way of modeling and visualizing a corpus of textual documents written in an unconstrained matter, or "free text", in a way that is intuitive for someone who is trying to discover semantic patterns in that corpus. Documents are often modeled using vectors of the words they contain, which can then be reduced in dimensionality, as in the common technique of latent semantic analysis (LSA). We aid the discovery of semantic correlations using less data by adding a semantic network of background common-sense knowledge, and analyze it in the same way. This is the core idea of AnalogySpace [10], a representation discussed further in the referenced paper, and this combined analysis allows us to provide the semantic models with more "intuition" [8].

The resulting dimensionality-reduced vectors are much easier to work with and compare to each other than the original data. Having expressed the data in a way that a computer can make some sense of it, however, we still desire an interface that represents this computational result in a way that a human can understand and work with it.

Luminoso

Luminoso[1] is an interactive application that aids a researcher in exploring these semantic spaces in a way that is intuitive for discovering semantic patterns from the dimensionality-reduced data. It enables them to create a vector space from a folder of input documents, using either an AnalogySpace-based model or a plain bag-of-words model, and then to explore that space interactively on a two-dimensional computer screen. The goal is to help the researcher understand their data by exploring this space, using an intuitive mouse operation we refer to as *grabbing*, which simultaneously lets them visualize the semantic neighborhood of the grabbed data point and use that point to N-dimensionally rotate their viewpoint. Other features of the interface help the user understand the space better, such as by using vectors with known semantics as "signposts".

Using Luminoso is a form of data mining that focuses on interactive exploration of the data. The importance of user participation in data mining has been observed by others [2], because an unsupervised algorithm to detect correlations in data will tend to find correlations that are spurious and irrelevant. An involved user, however, can guide this process toward relevant results by using their intuitive sense of what is interesting.

A use case for Luminoso that we focus on is to understand large quantities of people's suggestions and feedback at once. Survey forms frequently contain free-response spaces where people can write a paragraph to explain their views,

[1]This paper's content mirrors the paper in the main IUI conference entitled *Finding Your Way in a Multi-dimensional Semantic Space with Luminoso*.

IUI'10, February 7–10, 2010, Hong Kong, China.

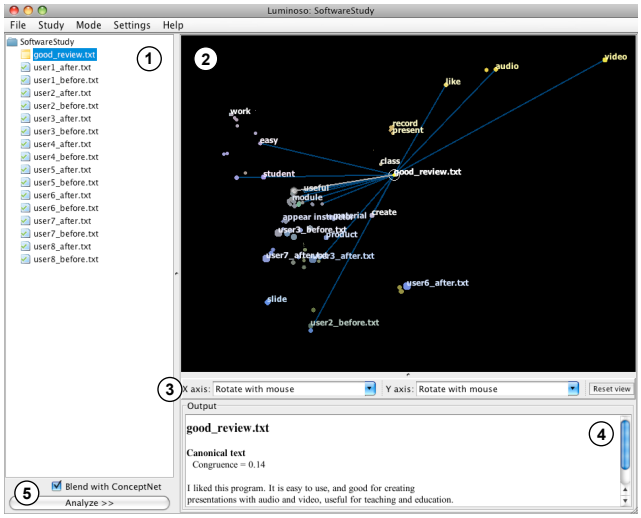


Figure 1. The overall interface to Luminoso, with labeled parts: (1) The document pane, allowing documents to be selected and new documents to be added to the study. (2) The viewer pane, providing a two-dimensional view into the SVD space. (3) The axis controls, with which the user can fix one or both of the axes to represent particular directions in the space. (4) The output pane, showing information about the selected point. (5) The "Analyze" button, which runs the SVD and updates the view. The "Blend with ConceptNet" option may be replaced with an interface for blending with any external data set in a future version.

but after a large enough number of people reply to such a survey, the free-response feedback tends to be ignored. Nobody has the time to read it all. By loading that data into Luminoso, however, one can visualize the major clusters of responses, view representative responses from each cluster, and even include known data about emotion or affect as signposts to understand the tone of the data.

CREATING THE SPACE

In order to display a representative space with Luminoso, we must first analyze the textual input used to create the space. We do this using a series of techniques designed to find patterns in natural language data – including patterns that appear within the input data, that come from a corpus of background knowledge, and that become apparent in the conjunction of both. We can use these techniques to draw general conclusions about the meaning of the data, cluster information in a variety of semantically informed ways, and make inferences across different types of information.

Natural language is a mode of input that can be handled particularly well by our techniques of common sense reasoning. We amplify the power of LSA, which is based only on the co-occurrence of words among the input documents, by including additional information about the semantic connections between words from ConceptNet. The additional knowledge this provides can help to better organize the words and phrases that appear in the input documents into a semantic space. It can recognize when two different words are semantically close to each other, such as "audio" and "video", even when this is not apparent from the distribution of word occurrences in the documents. It can also distinguish words that appear in different topic areas that exist independently of the input data, such as "action verbs", "household items", "computer terminology", and "things people don't want". This kind of information gives the vector space more power to represent the rough meaning of a document.

Applying common sense

ConceptNet [7] is a semantic network created using the information collected by the collaborative Open Mind Common Sense project. Using a representation that expresses knowledge as relations between words and short phrases, it describes the meanings of the words people use in terms of other words. The information contained in ConceptNet includes relations between everyday objects ("Books are used for reading."), information on people's priorities and goals ("People want to be respected."), and affectual information ("Arguments make people angry.").

AnalogySpace [10] refers to the technique of reasoning over such a semantic network by representing it as a matrix and performing singular value decomposition on it. Information in ConceptNet can easily be transformed into a matrix representation that relates its nodes (concepts such as "dog" or "taking pictures") to their neighboring edges (features such as "...has four legs" and "...is used for enjoyment"). Singular value decomposition expresses these concepts and features in terms of a core set of *axes*, or principal components, that are selected by the algorithm to represent the most variance in the data. The effect is to summarize the provided common-sense knowledge in terms of its large-scale patterns, using moderate-sized vectors (typically 50 to 100 dimensions) to represent each concept and each feature.

INTERACTING WITH LUMINOSO

The first step in interacting with Luminoso is to load the input documents. The container that holds documents and their analysis is called a *study*. The user can use the document tree to add documents to analyze (or they can use their operating system to drop documents into the folder representing the study). One or more of these documents can be marked as "canonical", which highlights it in the tree and makes it stand out in the interface, with effects that will be described later.

The user can choose whether to blend the data with ConceptNet in order to provide background information about semantics. Once the input is set up, the "Analyze" button creates the blend (if necessary), performs the SVD, and displays the results in the viewer window.

Congruence

A common use for a canonical document is to test whether the input documents generally "agree" with it semantically. As a way of assisting experimentation, the interface presents a statistic called *congruence* in the info pane when a canonical document point is grabbed. Congruence measures how much that canonical document aligns with the other documents in the study, which can also be seen as describing whether that document is typical or atypical among the input

data. This value can be compared between different runs of Luminoso or between different canonical documents.

The congruence of a document is calculated by comparing the distribution of cosine similarities between that document and all others, with the distribution of cosine similarities between all pairs of documents. The congruence is expressed as a Z-value (the difference in means over the standard error), so that it is scale-free.

GRAPHICALLY REPRESENTING N-DIMENSIONAL DATA

After using SVD to describe the data according to its principal components, one is left with vectors with a moderate number of dimensions. At this point, it is Luminoso's job to present this data understandably on a two-dimensional computer screen, so that the researcher can explore the resulting space, see whether it captures the patterns in the input data that it was intended to capture, and discover new patterns along the way.

The data can be represented as a sort of N-dimensional scatter plot. Each word, phrase, common-sense feature, or document in the input corresponds to a point in this space, which will use Luminoso to explore.

At any given time, Luminoso will project all the points in the N-dimensional space onto a two-dimensional plane, which the user can see a part of in a window on their computer screen. The user can change their viewport into this plane much like they would change their viewport in another 2-D interface such as Google Maps: the user can pan by dragging the right mouse button, or zoom using the mouse wheel or a laptop's equivalent "scrolling" gesture.

We represent each point as a small circle, at the appropriate location in the 2-D projection. The size of each circle increases with the number of times the item appears in the input, in order to draw attention to more significant inputs. Every point has a text label, describing a concept, a common-sense feature, or the name of a document, but not all of these labels can be displayed at once – the result would be incomprehensible clutter as many thousands of labels competed for screen space. Instead, only a subset of the labels are shown, determined interactively using the mouse pointer. The labels are chosen so as to set a maximum on the density of labels per unit of screen space. Additional points in a "full" area of the screen go unlabeled. The maximum density of labels decreases with the square of the distance from the mouse pointer. The effect is that, if the user wants to see the label of a point that is currently unlabeled, they can do so by moving the mouse closer to it.

Pressing the left mouse button will select the nearest point and "grab" it, which makes a number of useful things happen, one of which is that the user can use the grabbed point as a handle with which they can transform their view of the N-dimensional space. When the user grabs and drags a point, the view transforms (by stretching and rotating) in such a way that the point's projection onto the screen follows the mouse pointer, while the origin stays in the same place.

Related Work

Duffin and Barrett [4] describe an interface for rotating a projection, which differs from ours in that the user rotates the space by clicking and dragging a representation of an axis, instead of by clicking and dragging points in the space.

Buja et al. [1] describe the theory of projecting N-dimensional data onto a 2-dimensional view. This paper is largely concerned with creating "tours" of the space, or animations that trace a path between all possible projections of the N axes, but also mentions the ability to rotate particular axes using the "spider" interface.

Buja et al.'s paper provides a survey of existing software for multi-dimensional visualizations, such as GGobi [3], which performs singular value decomposition and allows visualizing the space using 2-D tours and spiders. Oelke et al [9] created a technique to understand and visualize customer data using opinion analysis using various techniques combined with clustering. This interface, unlike ours, is tuned specifically to understanding opinions in customer reviews.

Another interface that visualizes a corpus of textual data is described by Fortuna et al. [5] This interface displays LSA results with a labeled point for each term. Instead of using a direct projection onto the 2-D space of the screen, it instead uses multidimensional scaling to map the LSA space non-linearly into two dimensions. In Spire [11] has explored visualizing text using PCA; it focuses on the interactivity of their visualization space. However, unlike Luminoso, In Spire cannot take advantage of domain specific knowledge such as an ontology or work with NLP, such as bigrams [11].

APPLICATIONS

Increasingly, the commercial world has become interested in computational linguistics as a way to solve the problem of understanding customer feedback. Focus groups, consumer surveys, and other opportunities to communicate with customers often involve understanding their spoken or written text and "reading between the lines" to understand the patterns. Thus blending common sense with customer-generated free text can often yield insights that normal statistics miss.

The OMCS project worked with a large software company to analyze the data from their user tests [6]. In addition to rating various aspects of the software on a scale from 1 to 7, the users provided short-answer responses to various questions about their perception of the software. This free text data was considerably more expressive and informative than the numeric ratings, but the data was difficult to analyze automatically by computer. Blending with ConceptNet and exploring the results using Luminoso helped to draw general conclusions from the sparse data contained in the free text.

OPINION MINING: A CASE STUDY

To show how Luminoso can help with text understanding and opinion mining, we present a small example of using Luminoso on a data set of customer reviews. We looked up the reviews on Yelp.com for Thailand Cafe, a small restau-

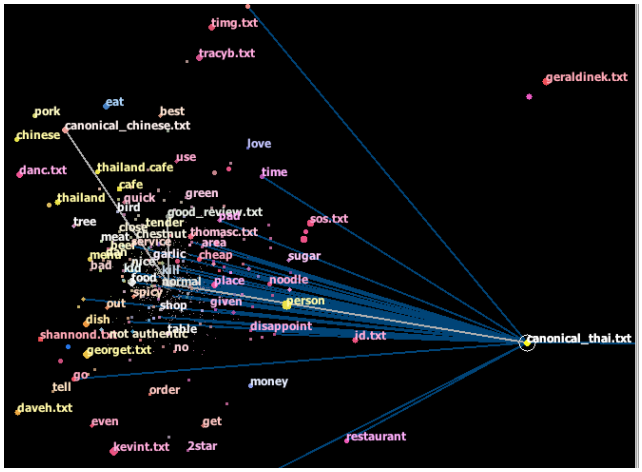


Figure 2. A Luminoso view showing reviews of a Thai and Szechuan restaurant. In this view, the top-to-bottom direction represents good to bad reviews, and the left-to-right direction represents a focus on Chinese cuisine versus Thai cuisine.

rant near MIT that is known for its inauthentic Thai food but is still popular with students. The restaurant recently started offering Szechuan food as well, and some people have observed that the Szechuan menu is better than the Thai menu. We can use Luminoso to explore how the new menu is reflected in customers' opinions of the restaurant. We would be able to see, for example, concepts associated with negative or positive restaurant reviews.

We made a Luminoso study out of the forty most recent Yelp reviews of Thailand Cafe (as of December 8, 2009), including the special symbols "1star" through "5star" to represent the star rating of each review.

We also added three canonical documents for the purpose of comparison: a positive review of another Thai restaurant, a positive review of another Szechuan Chinese restaurant, and finally, a synthetic document simply containing the text "5star 4star, not 2star 1star", which we use to mark the direction representing good reviews of Thailand Cafe.

The first two canonical documents are intended to show the difference between reviewers who are looking for Chinese food and reviewers who are looking for Thai. A view of the resulting space appears in Figure 2, where the synthetic "good review" vector is fixed to be the positive Y-axis.

The congruence values, as described on page 2, were 2.01 for the canonical Thai review, and -6.32 for the canonical Chinese review. We conclude that it was typical for reviewers to review it as a Thai restaurant (quite expected given the name), and atypical for reviewers to review it as a Chinese restaurant.

Meanwhile, from the positions of these canonical documents on the "good" axis (the Y-axis in Figure 2), we discover that there was a large correlation between the good reviews and the ones that reviewed it as a Chinese restaurant. No such correlation existed for the reviews that reviewed it as a Thai restaurant. In short, the people with a net favorable impression of Thailand Cafe were those who talked about its Chinese food, a promising result for their new menu.

REFERENCES

1. A. Buja, D. Cook, D. Asimov, and C. Hurley. Computational methods for high-dimensional rotations in data visualization. In C. R. Rao, editor, *Handbook of statistics: Data mining and data visualization*, pages 391 – 415. Lavoisier, April 2005.

2. A. Ceglar, J. F. Roddick, and P. Calder. Guiding knowledge discovery through interactive data mining. In *Managing data mining technologies in organizations: techniques and applications*, pages 45–87, Hershey, PA, USA, 2003. IGI Publishing.

3. D. Cook and D. F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*. Springer, December 2007.

4. K. L. Duffin and W. A. Barrett. Spiders: A new user interface for rotation and visualization of n-dimensional point sets. In *In Proceedings of the Conference on Visualization*, pages 205–211. IEEE Computer Society Press, 1994.

5. B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of text document corpus. *Informatica (Slovenia)*, 29(4):497–504, 2005.

6. C. Havasi. *Discovering Semantic Relations Using Singular Value Decomposition Based Techniques*. PhD thesis, Brandeis University, June 2009.

7. C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September 2007.

8. C. Havasi, R. Speer, J. Pustejovsky, and H. Lieberman. Digital intuition: Applying common sense using dimensionality reduction. *IEEE Intelligent Systems*, July 2009.

9. D. Oelke, M. Hao, C. Rohrdantz, D. Keim, U. Dayal, L.-E. Haug, and H. Janetzko. Visual opinion analysis of customer feedback data. In *In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2009.

10. R. Speer, C. Havasi, and H. Lieberman. AnalogySpace: Reducing the dimensionality of common sense knowledge. *Proceedings of AAAI 2008*, October 2008.

11. P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas. In-spire infovis 2004 contest entry. In *In Posters Compendium of InfoVis 2004*, pages 51–52, 2004.

User Analysis and Visualization from a Semantic Blog System

Jeong-Woo Son, Yong-Jin Han, Tae-Gil Noh, Seong-Bae Park, Se-Young Park
Department of Computer Engineering
Kyungpook National University
Daegu 702-701, Korea
[jwson, yjhan, tgnoh, sbpark, sypark]@sejong.knu.ac.kr

ABSTRACT
This paper describes a blog system which analyzes its users by extracting information from blog posts and visualizes the analyzed results. In this system, the events appeared in the blog posts are regarded as the information on bloggers and are stored in the *event ontology*. When a user submits queries about people, the information on bloggers is extracted from the ontology. It is used to find the paragons of the blogger communities, and then the user is compared with the paragons. The system presents the comparison results with four visualization components. A geographical visualization component shows both the volumes of the communities and the distances between the compared user and the paragons. An ego-central network presents actual connections between the user and the community members. A detail information chart enumerates the properties used in the comparisons in detail. Finally, a trend chart illustrates their change in time.

Author Keywords
Blogger community, Clustering, User Analysis, Visualization

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous

INTRODUCTION
The user-generated contents such as blog postings now have more traffics than the traditional main stream web sites [**?**]. Various kinds of bloggers' experiences such as book reading, making a trip, and so on are contained in the blog postings. Thus, blog postings can be regarded as the information on their writers. Since most legacy blogosphere can not handle all the specific information within the blog postings, their services for blogger analysis is limited to simple statistical information. If all the specific information is available, many interesting analyses of the bloggers are possible as follows.

- What are characteristics of the twenties in the west coast area?
- What are the notable differences between me and the bloggers in China?
- What are the characteristics of persons who are similar to me?

In order to obtain these kinds of knowledge, a system should handle not only the information within a blog post but also the information which emerges from the collections of postings. This type of knowledge is called as Emergent Knowledge [**?**]. A collective knowledge system which handles the emergent knowledge is defined as "the system which enables computation and inference over the collected information, leading to answers, discoveries, or other results that are not found in the human contributions." A travel social web system with a recommendation facility was given as an example of such systems in [**?**]. The similar efforts of processing and finding collective knowledge can be found in [**?**] (a recommendation system) and [**?**] (a tutor system).

This paper proposes a blog system which analyzes its users with the information within blog postings. This system assumes that the emergent knowledge about bloggers is obtained by revealing blogger communities. To obtain such knowledge, blog postings are first stored in the *event ontology* [**?**] which is a machine understandable format after the postings are semantically annotated. The system regards the postings of a blogger as the unique information of the blogger. Therefore, when a user summits a query about people such as "the twenties, movie fan", the information on the bloggers is extracted from the ontology. To know the similarity and difference between the user who submitted the query and the bloggers who meet the query, the bloggers are clustered into a few number of communities. The centroid of each community can be regarded as a paragon of the community. Then, the user is analyzed by comparing her with the paragons.

The system presents the comparison results with four visualization components. A geographical visualization component shows both the volumes of the communities and the distances between the user and the paragons. An ego-central network presents actual connections between the user and the community members. A detail information chart enumerates the properties used in the comparisons in detail. Finally, a trend chart illustrates their change in time.

Table 1. Properties and their value type for subject and object.

Category	Feature Name	Value Type
Subject	age	Numeric
	gender	Binary
	location	String
Object (Book)	genre	String
	authors	
	publisher	
Object (Trip)	address	
	type	
Object (IT device)	type	
	manufacture	

Table 2. Definition of distances for various value types.

Feature type	Distance
N: numeric	$Dist_N(x, Co) = \frac{\sqrt{(\sum_{i=1}^{k} Co_i - x_i)^2}}{max}$
B: binary	$Dist_B(x, Co) = \frac{\sum_{i=1}^{k} (I(Co_i, x_i))}{\sum_{i=1}^{k} Co_i}$
S: string	$Dist_S(x, Co) = \log(\frac{1}{Z_F(Co)^{\sum_{i=1}^{k} I(Co_i, x_i)}})$

INFORMATION EXTRACTION FROM EVENT ONTOLOGY

The event ontology [**?**] is designed to express simple events of everyday like such as dining, shopping, reading book, or making a trip. The ontology is used as a vocabulary to annotate each blog posting. To annotate blog postings, the proposed system uses the semantic blog system proposed by [**?**]. When a blog is posted, it is expressed in the ontology with three major instances of a subject, an object, and an event. The subject is an instance of FOAF ontology and contains the information of the blogger including age, gender, job, and so on. The object describes what the blog post is about, and it is an instance of a domain ontology. The event represents the type of the post such as book review, listening music, or so on. A number of domain ontologies [**?**, **?**, **?**] are used to capture various event types of the blog posts.

When the information of each blogger is extracted from the event ontology, all properties of subject, object, and event are used as features of the blogger. Event has a single feature which is the number of posts for its post type. For subject and object, all properties shown in Table 1 are used as features.

USER ANALYSIS BY CLUSTERING BLOGGERS

When a user submits a query, the system selects bloggers who meet the query and extracts their information. After that, the user is characterized by being compared with the selected bloggers. However, such comparison could be meaningless as the number of the selected bloggers increases, since the large community has a tendency to be general or characterless. In the proposed system, for the comparisons of the user with blogger communities which have a specific characteristics, a number of clusters of the selected bloggers are first found. These clusters of bloggers are considered as blogger communities and the comparison of the user with the blogger communities is done actually by comparing the user with the paragons of all blogger communities.

In order to cluster the bloggers, a distance metric is needed which measures the similarity among bloggers. In designing a distance, various types of feature values should be considered, since ontologies have various value types for their properties.

Let x be the user who summitted a query and Co be a paragon in a community. Then the distance between x and Co is defined as

$$\begin{aligned} Dist(x, Co) &= w_N Dist_N(x, Co) \\ &+ w_S Dist_S(x, Co) \\ &+ w_B Dist_B(x, Co), \end{aligned}$$

where N, S,and B are feature types described in Table 1 and w_a is a weight for the feature type a and is defined as

$$w_a = \frac{\text{No. of features for type } a}{\text{No. of all features}}.$$

The definition of each distance function is shown in Table 2. In this table, k denotes the number of users who belong to the community Co, and $I()$ is an indicator function. Here, Co_i is the value of the i-th feature of Co. Function $Z_F(Co)$ returns the number of features whose value type is F. To reflect characteristics of each feature type, the distance function, $Dist_a(x, Co)$, is defined according to the feature type a. An Eucleadian distance is adopted for the numeric type. Since max is the maximum of all Euclidean distances, this distance is the one normalized. For the binary type feature used to express gender of users, the empirical probability that the same gender appears both at x and Co is used as a distance. Finally, the string type regards, as a distance, the number of feature values both x and Co are shared.

With this distance, bloggers are clustered to form communities. For this purpose, k-means clustering [**?**] and hierarchical clustering [**?**] are adopted. First, k-means clustering or hierarchical clustering is performed to get blogger communities. Then, the centroid of each cluster is set as a paragon of the cluster which represents a community of bloggers.

USER ORIENTED ANALYSIS AND VISUALIZATION

In the proposed system, the user who submitted a query is analyzed by being compared with blogger communities and their paragons and the comparison result is illustrated using four visualization components. These visualization components are used to view the comparison result in different point of view.

First of all, the difference and the similarity between the user and each paragon are presented in the geographical visualization component as shown in Figure 1. The left side of this figure shows personal information of the user such

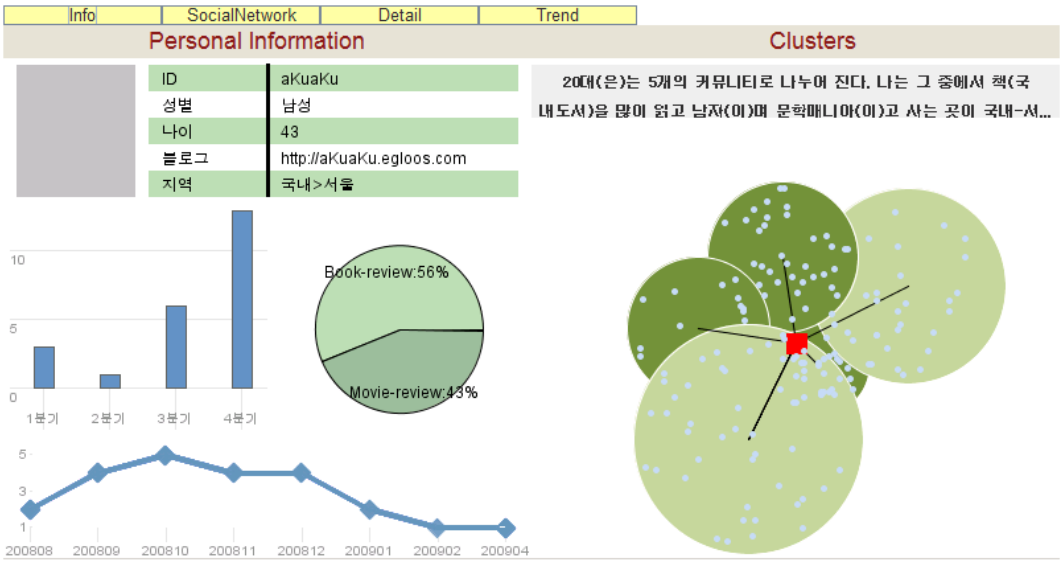


Figure 1. Clustering component.

as ID, gender, blog URL, and so on. The graphs on this side present simple statistical information of user postings. In the right side, the central node denotes the user and the cylinders represent blogger communities. The radius of a cylinder is determined with the mean distance among community members and its height is set using the number of community members. Dots on a cylinder shows the number of community members and their density. For easily explaining communities to the system users, texts explaining the paragons are also given in this component when the cursor is over wrapped . Distances of the user against each paragon are represented using edges. The label on right upper side describes information of the paragon in the most similar community.

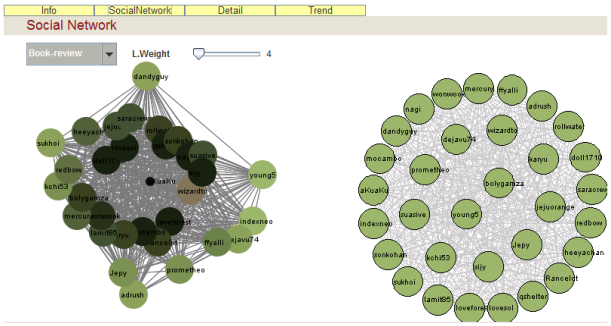


Figure 2. Ego-central network component.

The next visualization component shows two kinds of networks between the user and community members. Figure 2 shows this component. The network on left side shows ego-centric network which the central node is the user. The length of edges to each blogger is the distance from the user to the blogger. Using the listbox on this network, the user can choose the kind of edges such as 'movie', 'book author', 'actor' and so on. The color of each blogger node is determined using the similarity between the blogger and the user. Dark color means more similar blogger than nodes with bright color. In the right side, ordinary social network is presented. For both networks, the size of the blogger nodes is determined using betweenness, a kind of centrality [**?**] which denotes an influence of bloggers in a community.

Even though the connections to blogger communities and their members are presented using the two components, it does not present the detailed difference according to the features explained above. For showing the difference and the similarity according to each feature, the detail information chart shown in Figure **??** is adopted. In this figure, the chart compares the user and the communities by the quantity of feature values, where the green bar is drawn using the information of bloggers while the blue bar is drawn using that of the user. The detail information chart is composed of three charts. Each chart gives the information of bloggers and the user according to object types of book, IT device, and movie.

The last component is a trend chart shown in Figure **??**. This component represents a trend of the user by comparing her with a community. The X-axis is time and Y-axis implies the frequency of the feature value in the given time. The user can choose a specific feature value or a set of feature values to see a trend using the list box at the top of this component. In this figure, it is shown the trend of the user and a community for korean books.

In case of the detail information chart and the trend chart, the objects appeared in postings of the user or bloggers are listed using the object list component shown in Figure **??**. In this component, the objects can be sorted according to their frequency or annotated rate and some objects appeared in both the user's and bloggers' lists can be filtered using the radio button.

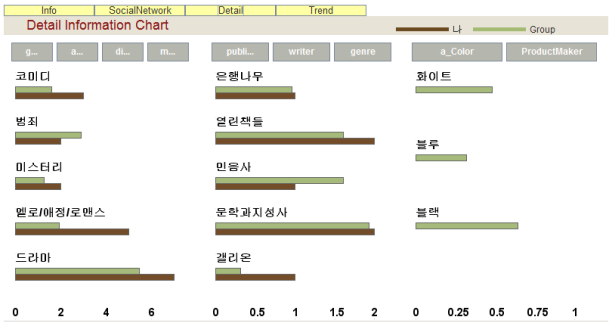


Figure 3. Detail information component.

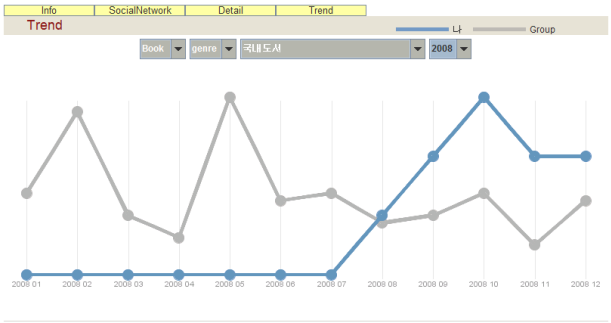


Figure 4. Trend component.

CONCLUSION

This paper described a blog system which analyzes its users with the information within the blog postings. In the system, the user who submitted a query is compared with the bloggers who meet the query. To compare the user and bloggers with specific characteristics, the system clustered the bloggers into several communities. After then, the comparison is done using paragons of the communities.

The comparison result is presented with four visualization components. The geographical visualization component and the ego-central network present distances between the user and paragons or community members. The detail information chart enumerates the features used in the comparisons in detail. Finally, the trend chart illustrates their change in time.

Acknowledgements

This work was supported in part by MIC & IITA through IT Leading R&D Support Project and by MIC & IITA through IT Leading R&D Support Project (A1100-0601-0102).

REFERENCES

1. Technorati report: State of the Blogosphere 2008. http://technorati.com/blogging/state-of-the-blogosphere.
2. C. Bizer, R. Cyganiak, and T. Gauss. The RDF Book Mashup: From Web Apis to a Web of Data. In *Proceedings of the 3rd Workshop on Scripting for the Semantic Web, 4th European Semantic Web Conference*, pages 389–393, 2007.

Figure 5. Object List component.

3. P. Carrington, J. Scott, and S. Wasserman. *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005.
4. K. S. Choi. IT Ontology and Semantic Technology. In *Proceedings of Natural Language Processing and Knowledge Engineering 2007*, pages 14–15, 2007.
5. T. Gruber. Collective Knowledge System: Where the Social Web Meedts the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):4–13, 2008.
6. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.
7. D. Mackay. *Information Theory Inference and Learning Algorithms*. Cambridge University Press, 2003.
8. T. Noh, Y. Han, J. Son, H. Song, H. Yoon, J. Lee, S. Lee, K. Kim, Y. Lee, S. B. Park, S. Y. Park, and S. J. Lee. Experience Search: Accessing the Emergent Knowledge from Annotated Blog Postings. In *Proceedings of the Symposium on Social Computing Applications, The 2009 IEEE International Conference on Social Computing*, 2009.
9. C. Torniai, J. Jovanovic, D. Gasevic, S. Bateman, and M. Hatala. E-Learning Meets the Social Semantic Web. In *Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies*, pages 389–393, 2008.
10. Y. Wang, N. Stash, L. Aroyo, P. Gorgels, L. Rutledge, and G. Schreiber. Recommendations Based on Semantically Enriched Museum Collections. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):283–290, 2008.
11. M. Wick, B. Vatant, and B. Christophe. Geonames Ontology. http://www.geonames.org/ontology/.

Integrating Interactivity into Visualising Sentiment Analysis of Blogs

Hyowon Lee, Paul Ferguson, Neil O'Hare, Cathal Gurrin and Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland
hlee@computing.dcu.ie

ABSTRACT
With an increased amount of freely available online resources and strong interest in automatic crawling and analysis on such resources, suitable visualisation techniques to present the results of such analysis present an important agenda for the visualisation research community. Interactivity on the Web has also become much more commonplace and acceptable as today's Web technologies such as Web 2.0 and Flash become more widespread. While conventional graphs and charts augmented with interactivity are one way to present the output of analysis, an interaction strategy that leverages the interactivity style of the Web should be more suitable than what we see today. We present a novel interactive visualisation technique designed and implemented on top of text-based sentiment analysis for financial blog posts where a user can easily search and browse bloggers' aggregated opinions on commercial companies in a way that helps understand the levels of online opinion in a summarised as well as a detailed manner.

Author Keywords
Interactive Visualisation, Sentiment Analysis, Financial Blogs

INTRODUCTION
Studies on crawling Websites such as blogs, news articles, product reviews and political columns and then automatically extracting useful information or determining their meaning, has become an increasingly active area of research. This helps realise the great potential for leveraging the rich online resources available today. Sentiment Analysis of blog posts is one of these efforts: trying to determine bloggers' opinions in terms of positive or negative sentiment, by analysing the text contained in the blog posts. As with any other similar text analyses, an important issue derived from such an analysis is how to present the results to users in a way that facilitates easy understanding of the overall trend of blog sentiment, as well as specific instances of such blogs.

The question we ask (and the solution we present) in this paper is how we could leverage Web interactivity in visualising the sentiment of blogs as a result of crawling and analysing a large number of such blogs. Conventional charts and graphs are an effective static visualisation tool but how can we better incorporate the dynamic and higher levels of interactivity that suit the style of today's Web interaction ?

Considering the increasing interactivity on today's Web services and its' popularity, leveraging the style of Web interactivity and exploring visualisation strategies in that direction is well worth an investigation. In this paper we present a novel interactive visualisation strategy and the resultant Web interface allowing its users to interactively search and browse the output of sentiment analysis, similar to the way Web users search and browse Web pages using a Web search service.

VISUALISING SENTIMENT ANALYSIS — ALTERNATIVES TO PIE CHARTS AND BAR GRAPHS ?
Visualising the blogosphere has become an increasingly popular research challenge, and a number of graphical representations have been adapted to visualise the results of blog text analysis, although most of these are still at an experimental or planned, rather than deployed, stage. Examples include the use of different sizes of rectangular areas and colours as positive/negative opinion indicators [4], [1], spatial segmentation of Google News stories with colour-coded story categorisation[1], a pie-chart-like petal visualisation of different facets of sentiment such as Positive/Negative, Cooperative/Conflict, Pleasure/Pain and Virtue/Vice [6], the use of a large area to plot months of US presidential election events by colour-coding between Republican (Red) and Democrat (Blue) [13], multiple mini bar charts to visualise different facets of home electronics such as LCD, battery and speaker [11], blog visualisations where positive/negative emotions are colour-coded in blue/red and presented as a stack of horizontal bars [5], and where blog entry length, comment length, and the number of posts by the same bloggers are mapped to visual properties (colour, circle size and distance from timeline) [10]. More conventional graph/chart-based timeline visualisations of consumer-generated data include MoodViews [3], ThemeRiver [8], and a method that uses a Time Series Data Processing technique [2].

Other innovative blog visualisations include Twingly Blog-

[1]Newsmap: http://marumushi.com/apps/newsmap/index.cfm

CHI 2009, April 3 - 9, 2009, Boston, MA, USA.

stream[2] which analyses and visualises the linkage status of bloggers in a time graph, where such linkage status amongst blog messages can be visualised as 3-D shapes [9]. "We Feel Fine" [7] crawls blog websites and identifies where a blogger's feelings are mentioned, and visualises these as animated keyword clouds with a montage of images taken from the blog sites on a given day. We can imagine how these could be extended to the style of well-known 3-D visualisation techniques such as Cone-Trees, Document Lens and Perspective Wall where large, high-resolution monitors are usually assumed. In this paper we address the question of what would be an effective visualisation strategy, other than variations of conventional graphs and charts above, which is at the same time Web-friendly yet integrates even more user interactivity.

INTEGRATING INTERACTIVITY INTO SENTIMENT VISUALISATION

In this section, we briefly describe the sentiment analysis system we have developed, then present the novel interaction strategy and detailed design considerations and decisions we have taken to realise the strategy.

System for Sentiment Analysis on Financial Blogs

The sentiment analysis system at the back-end of the interface described in this paper was developed from a collaboration between Dublin City University and Zignals[3], a company working in online stock trading. The aim of the system is to automatically extract subjective opinions found on blogs and to track the changing sentiment from the blogosphere towards individual stocks and the market in general. The system has been crawling financial weblogs from over 170 sources since May 2009, and has to date crawled over 44,000 relevant articles, namely those relevant to any company in the S&P 500 list (currently our system has analysed over 34,000 article-company matches). These are then analysed for sentiment (positive, neutral, negative) towards that company, using topic-based sentiment analysis approaches described in [12]. The results of this sentiment analysis is then aggregated for the interactive visualisation described in this paper.

Unit Representation

The core of our interactive visualisation is the concept of "Unit Representation", a visual representation of the result of sentiment analysis on a particular object (a company in our case), serving as the building block of the overall interaction our visualisation uses for searching and browsing.

Figure 1. Unit representation - defining a visual representation for interactive querying and browsing

[2]Twingly Blogstream: http://www.twingly.com/enterprise
[3]Zignals: http://www.zignals.com

Figure 1 shows such a Unit Representation depicting (1) the name and logo of a company (VF Corp in this case), (2) the number of blog sources used for analysis (24 articles from 7 Websites in this case) and (3) the aggregated level of opinion (+2 on a scale between -3 and +3, and amounting to 3.5% increase compared to a previous time period in this case).

Searching and Browsing

Once the Unit Representation is defined as in Figure 1, then it can be used as a 'virtual document surrogate' that can be presented as a unit of retrieval on the user-interface. In other words, a user can conduct a search and the result is a list of Unit Representations, ranked initially by the order of the analysed level of opinions. The user can further select an entry in the search result, browse more details, check where the sources of this level of opinion came from, etc. Figure 2 shows a screenshot of the overall system interface.

On the top left of Figure 2, a user starts by selecting a category of companies (Finance, Medicine, Insurance, etc. currently Technology is selected in the figure) and the overall sentiment on all companies in the selected category is indicated on the right of the category selection. Below this, the user can further specify a time interval by clicking on mini calendar icons and selecting a date, or by clicking on common time interval types (week, month, and year) upon which 'from/to' boxes will be adjusted to the selected interval. When the user clicks on the 'GO' button, the result will be presented below, as a list of Unit Representations. The search result can be sorted by the levels of positive/negative sentiment, the rate of opinion change, company name, or the number of articles that mentioned the company, by clicking on the sorting buttons (below the 'GO' button). Clicking on any of the entries then presents all the articles that were used to derive the opinion rating in the Articles panel in the middle of the screen (in Figure 2 the user selected Microsoft and the Articles panel presents the articles that refer to Microsoft). At the top of this panel a summary of the opinions from all articles that refer to the selected company is presented, and below it is a list of articles in a summarised format. Each article entry also shows whether it has a positive, negative or neutral opinion about the company with a small colour circle beside the article's title. The article entries can be sorted by the level of positiveness/negativeness, title of the article or by the date of post, similar to the sorting feature of the search result panel on the left. Clicking on the title of an article then opens up a web browser window and brings the user to the original online article so that the user can read the full article from the source Website. Finally on the right side of the screen is the Sources panel that shows a list of source Websites and the number of articles used in the analysis of the selected company, changing as the user selects a different company.

As can be seen in this interaction, the strategy employed turns the conventional concept of static graphical representation (which most financial information and company profile charts use) into an inherently interactive, search-like interface where the user starts with querying followed by sorting the search result to see the results in different orders and then

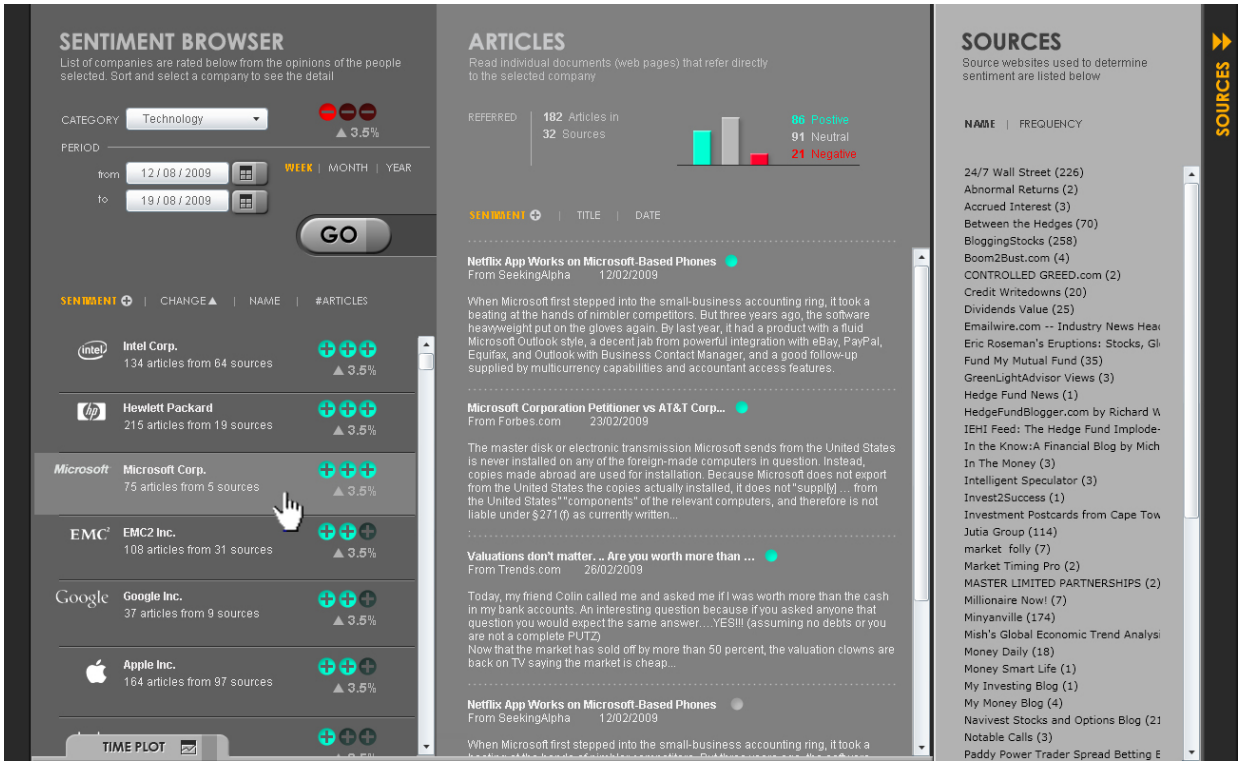


Figure 2. Putting together - stacking Unit Representation as a result of searching

browsing for more detail on the entries in the results.

Opinion Changes Over Time

As changes of opinion over time can be effectively presented and easily understood with a conventional time-based graph, such a graph can be incorporated into our overall design simply as a panel at the bottom of the screen which can slide up or down as the user wishes. If a user wants to view the temporal changes of an opinion about a particular company, s/he selects a company on the search result panel and drags it down to the bottom of the screen where the Time Plot panel tab is located (see Figure 2). This will then slide up the Time Plot panel and present the selected company's profile change over the user-specified period (see Figure 3). The user can drag in more than one company into this slide-up panel to compare the profile changes of multiple companies.

In Figure 3 the user has dragged two companies (Intel Corp. and Hewlett Packard) into the Time Plot panel, and each is assigned a unique colour (orange and red). Bringing the mouse cursor on the entry on the left of the Time Plot panel highlights the line in the graph area on the right, with vertical dotted lines at each of the data analysis points indicating the variance of opinions at that point in time. In Figure 3, the Intel Corp. time plot shows that the opinions improved over the past 1 week with the variance of opinions decreasing (i.e. opinions converging) as the orange line and its vertical dotted lines indicate. At any time the user can click on the Time Plot panel heading to slide it down or up, trading off the area with the list of companies presented above. On a very large computer screen we could facilitate a permanent area for Time Plot panel without sacrificing other information areas, but in the current implementation we assumed a computer monitor in a typical office, and thus we adopted such a slide-in and -out panel solution.

CONCLUSION

Facilitating interactivity in visualisation does not necessarily mean conventional graphs and charts augmented with a few animated or mouse-over effects. Our contribution in this paper is to explore and introduce a new interactive visualisation where an individual retrieval unit is visually defined then it is used as the unit of searching and browsing as if one searches Web pages, rather than attempting a more conventional high-density visualisation schemes often tried in the visualisation community. Thus the issue here is not so much on how much information density one screen can accommodate (e.g. how many companies the left panel can display) similar to the fact that a scalability is not an issue on most Web search engines' search result display where only top few entries are of concern to the user and he/she can easily sort or filter the order of entries. The system is fully implemented with its Web interface running on Silverlight. While we have had a series of informal user tests during the prototype development stage with the interface, we are now planning to conduct a more formal evaluation with the complete system in order to gain a better understanding of the ways this strategy can support a specific set of tasks in the financial domain.

ACKNOWLEDGMENT

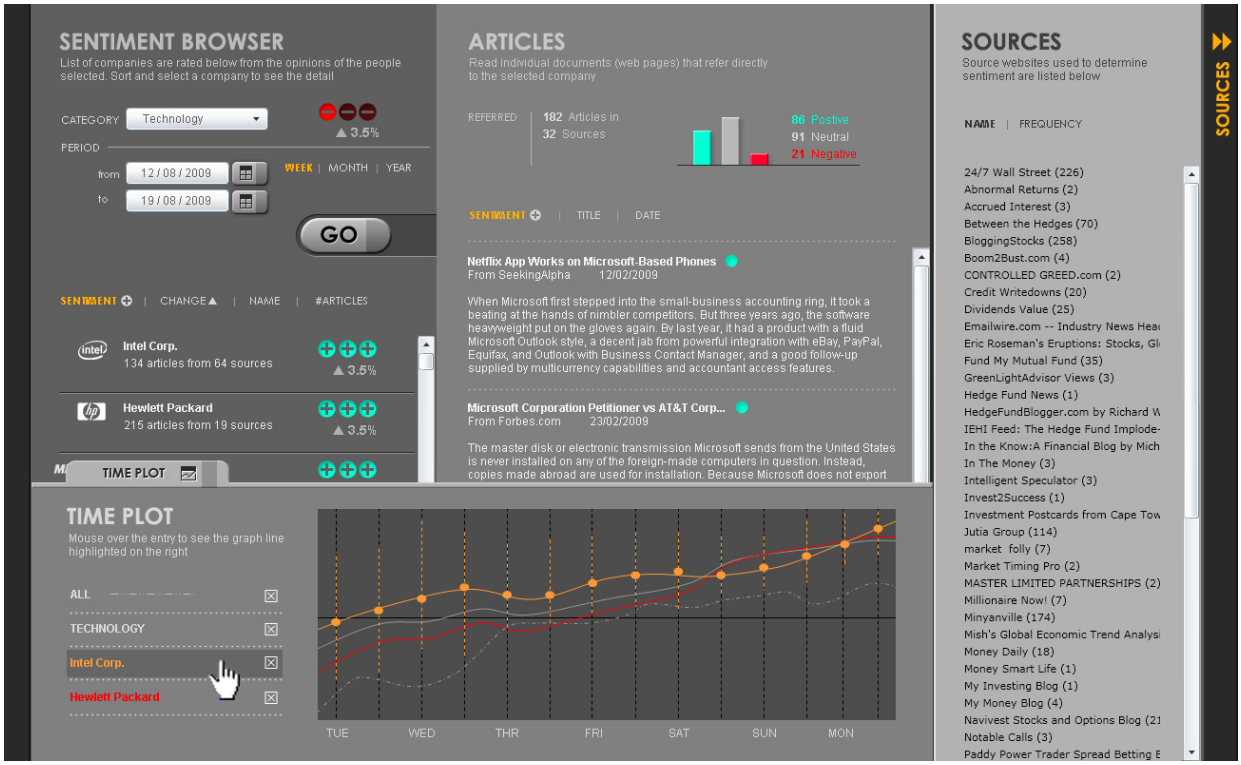


Figure 3. Time Plot - looking at the opinion changes over time and comparing those for multiple companies

This work is supported by Science Foundation Ireland under grant 07/CE/I1147, and by Enterprise Ireland under grant IP/2008/0549.

REFERENCES

1. G. Carenini, R. T. Ng, and A. Pauls. Interactive multimedia summaries of evaluative text. In *IUI '06*, pages 124–131, New York, NY, USA, 2006. ACM.
2. T.-c. Fu, D. C. M. Sze, P. K. C. Leung, K.-y. Hung, and F.-l. Chung. Analysis and visualization of time series data from consumer-generated media and news archives. In *WI-IATW '07*, pages 259–262, 2007.
3. T. Fukuhara, H. Nakagawa, and T. Nishida. Understanding sentiment of people from news articles: temporal sentiment analysis of social events. In *ICWSM 2007*, 2007.
4. M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, pages 121–132, Berlin/Heidelberg, 2005. Springer.
5. M. Gamon, S. Basu, D. Belenko, D. Fisher, and M. Hurst. Blews: Using blogs to provide context for news articles. In *ICWSM 2008*. Association for the Advancement of Artificial Intelligence, 2008.
6. M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner. User-directed sentiment analysis: visualizing the affective content of documents. In *SST '06*, pages 23–30, 2006.
7. J. Harris and S. Kamvar. We feel fine. an exploration of human emotion, in six movements. available at: http://www.wefeelfine.org/, 2009.
8. S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.
9. M. Hurst. Data mining: Mapping the blogosphere, from text minding, visualization and social media, 2009. http://datamining.typepad.com/gallery/blog-map-gallery.html.
10. Indratmo, J. Vassileva, and C. Gutwin. Exploring blog archives with interactive visualization. In *AVI '08*, pages 39–46, New York, NY, USA, 2008. ACM.
11. B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05*, pages 342–351, New York, NY, 2005.
12. N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, and A. F. Smeaton. Topic-dependent sentiment analysis of financial blogs. In *TSA'09*, Nov 2009.
13. F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. Keim. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *VISSW 2009*, February 2009.

A Visual Approach to Text Corpora Comparison

Dixon Ip, Ken Lau Ka Keung, Weiwei Cui, Huamin Qu and Helen Shen
{dixonip, kenlau, weiwei, huamin, helens} @cse.ust.hk
The Hong Kong University of Science & Technology

ABSTRACT

Email is an important part of our lives, almost everyone has at least one email account. It is not uncommon for an average professional email user to receive/send over a hundred emails a day. With all these personal and professional messages accumulated over time, email becomes our personal archive. In recently years, many interesting techniques are developed to improve email management, visualize mailbox content, explore historical events, and discover hidden knowledge from these electronic archives. In this paper, we propose an approach to visualize and compare email archives with Streamgraphs and Tag Clouds. Our preliminary results demonstrate the capability of this technique to compare any general text corpora with temporal information.

1. INTRODUCTION

Electronic mail (email) is a method of exchanging messages digitally. Since the first email was sent in 1971 via the ARPANET, email has becoming one of the most important communication tool widely adopted for academic, professional and social purposes. Today, information explosion is a problem faced by everyone bombarded by unstructured data from online news, forums, chat groups, instant messaging, especially emails. Email is an important part of our lives, almost everyone has at least one email account. Any advance in email systems impacts everyone. A person uses email to communicate informally with her/his friends and family, professional managers use email to dispatch or delegate tasks to their subordinates, and researchers collaborate with team members over a wide geographical regions and time zones using email as the media. With all these personal and professional messages accumulated over a period of time, email becomes our personal archive. Email was invented as a communication tool, it has also become a burden to many email users receive hundreds of messages a day. Whittaker & Sidner [19] has coined the term "Email Overload" in 1996 as they have discovered email users overload the email system by using email for purposes other than simply messaging such as task management, contact management, calendaring and personal archiving. Ten years later, Fisher et al [5], revisited this Email Overload problem and concluded Email Overload will continue to grow as size of email archives had increased 10 times from 1996 to 2006; more email users use email as personal archives. Gemmell et al. [8] even advocate the need to develop a platform as better mean to capture "everything" of a personal's life time into a personal database. Before we have a new system designed for the purpose of personal archiving, email archives provide a "goldmine" for researchers and scientists to learn more about a person or an organization. There are previous research works related to methods to improve email management, many of these efforts fall into Computer-Human Interface design. With the increase in the interest of and advance of information visualization, many researchers attempt to visualize one's mailbox or email archives with conventional visualization techniques. In this paper, we discuss the motivation of email visualization and related works. At last, we conclude our discussion with future research opportunities.

2. MOTIVATION

In the commercial market, there are advanced email clients offer labeling/tagging and clustering of email threads for better management of inbox and visualization of conversations embedded in the mailbox. In addition, there are desktop search tools (e.g. Google Desktop) help to locate specific items in archives. The improve in computer-human interface and personal search tools both aim at allowing the users to better manage emails or locate very specific piece of information from large archives. However, users have to provide very specific search rules with a crystal clear understanding of what they are searching for. Moreover, they do not provide a holistic view of information contained in the archives. Scientists, historians and investigators see email archives as important artifacts for understanding the individuals or the organization; however, there is no promising method to effectively explore these archives. Visualization creates a new perspective as a potential solution. Preliminary results from email visualization demonstrate the capabilities to discover new knowledge based on communication patterns of people, topics/themes and temporal information. Techniques can be adapted for improvement of computer human interaction design in email clients and analysis of email archives for knowledge discovery purposes. In a more general topic in text visualization, there are many research efforts on contrasting documents. However, there is no previous attempt to visualize discussion trend or change of themes over time at the institutional level or any related work on comparing topics/themes between corpora with temporal information.

3. RELATED WORK

The techniques employed in email visualization are highly dependent on the data types of the attributes selected for visualization, we will discuss individual works in this section. Email researchers attempt to visualize either the email conversation with temporal information or the themes with semantic analysis on keywords by exploiting the social network

information and the communication patterns contained in the email archives. We see a need for a framework to study the different approaches. Perer et al., [11] illustrated a framework to subdivide types of interactions with current emails and archived emails. In this paper, we adopt this fundamental framework with a further subdivision by the purpose of the visualization projects. There are two major perspectives in the visualization: (1) individual's archives, and (2) institutional archives. Works at the individual-level, either focus on managing or exploring email archives of an individual. A few research works at the individual-level provides visuals for users to reflect on historical events. While research works at the institutional level mainly focus on providing a holistic view of the archives or creating a portal to support dynamic exploration of the archives.

Under the individual's perspective, a few attempt to improve email clients by integrating some simple visualization techniques. The Remail project [13] develops email arcs to visualize the email threads and provides a correspondents "heat" map to allow user to see correspondents in prioritized view when replying emails. SNARF [6] is another email client project, which focuses on unread mails prioritization also allow users to view email threads in an organized tree-like view. Both Remail and SNARF adopt semantic analysis to understand the email content in addition to the usual information of Sender, Recipient, Subject, Date/Time, etc in the message header. Besides, SNARF also study user behavior in replying emails to determine importance of correspondents. The next category under individual perspective consists of tools for users to reflect on their email actions and histories. Viégas et al. [17, 18] has a few email visualization projects. In his first project, Mountain provides a 1-to-many correspondents view based on the volume of email communication between an individual and her/his email correspondents over time. This is the first attempt on providing a single view of an individual's mailbox with temporal information. Layers of correspondents form the mountain, a new peak can indicate significant change in life, e.g. joining a new company associates with a sudden increase in number of new correspondents/layers in the mountain. Viégas also develops PostHistory [17] which depicts quantitative aspects of a user's email activity on a daily basis by providing a heat map like calendar-view alongside with histogram timeline to show the frequency of correspondents over time. Themail [18], also by Viégas highlights the overall patterns of communication with one correspondent at a time. Themail uses semantic analysis on the email content to show the words that characterize one-to-one correspondence over time. One can visualize the change in the theme of correspondence that may also signify change of relationship over time.

Moving onto tools for exploring individual's mailbox, Frau et al. [7] and Perer [12] have developed Mailview and Contrasting Portraits respectively. Both are based on conventional visualization techniques such as treemap, scatter plot and histogram Timeline to provide straightforward visuals on overall communication history. Mailview focuses on emails to correspondents relationship over time, whilst Perer demonstrates the use of classical techniques like: (1) treemap to visualize the organizational hierarchy of email correspondents, (2) scatter plot based on the number of messages sent to the correspondent against the number of messages received from the correspondent, and (3) histogram timeline to contrast number of incoming messages versus outgoing messages. Liu et al. [9] developed an interactive visual text analysis tool, called TIARA (Text Insight via Automated, Responsive Analysis). On the visual aspect, they utilize an augmented stack graph to provide a high-level visualization of a personal mailbox as an example, and they also demonstrated the possibility of visualizing general text corpora with the same approach over a timeline.

At the institutional level, we have found only one interesting visualization, Enron Explorer [16], developed by Trampoline Systems' attempt to provide a platform to support exploration. Enron Explorer is an online tool developed by Trampoline Systems after the company Enron went bankruptcy. Trampoline engineers used this Enron email dataset (200,000+ emails) as a test bed during development of the company's information mapping technology. The Enron Explorer lets you visualizes the overall communication pattern of the senior management of Enron. It analyses each person's main contacts and the themes they are talking about. This visualizer presents the social network within Enron. User can drill down to the mailbox of individual employees of Enron by following a specific theme. One major deficiency of this visualization is lack of temporal information in the representation.

In pure text visualization, the common efforts are on abstraction and summarization of articles and books with keywords like tag clouds, parallel tag clouds [4], and TextArc [10]. Recently there are more works on document contrasting. Clark [3] has published some interesting work on comparing text, speech or documents with Streamgraphs and Document Contrast Diagram. The comparison with Streamgraphs is a straight forward side-by-side comparison whereas the Document Contrast Diagram illustrates the shared and unique keywords in two documents in a scatter plot style. The horizontal position and color of the keywords reflects the frequency of occurrence in the documents. Their vertical positions tie to the time of appearance in the documents. The Document Contract Diagram incorporates temporal information very loosely as a user cannot tell the exact appearance of keywords. To display a closer relation of the keywords to the timeline, Thorp [15] has developed a tool for examining similarities and differences between any two articles. The idea is similar to Clark's, Thorp uses links to connect the keywords to the documents' timeline to illustrate very single occurrence. All of these visual techniques inspire our work on a bi-level approach to compare two text corpora at the holistic theme level and at the detailed keyword level. Another critical requirement is the capability to contract the two text corpora in respect to the timeline. The following section introduces our visual approach: Polarized Streamgraph.

4. POLARIZED STREAMGRAPH

There are three critical requirements for comparing text corpora with temporal information such as email archives, speech and news.

1. Contrasting: Theme/Topic detection, contrasting the similarities and differences.
2. Trending: Change of theme/topic over time, showing the trend.

3. Drilldown: Bi-leveling visualization, highlighting the difference at the theme-level and detailing the contrast at the keyword-level on demand.

Our approach accepts the output of any statistical methods such as term frequency–inverse document frequency (tf-idf) for keyword extraction [14] and Latent Dirichlet Allocation (LDA) for topic modeling [1]. In our paper, we focus on the discussion of the visualization techniques.

We adopted streamgraph as a natural visual tool to illustrate the trending of themes/topics over time. At shown in figure 1, each layer of the streamgraph represents an unique topic extracted from the two text corpora. The similarities and differences of the two corpora are contrasted by colors. In this color encoding scheme, blue and red represent corpus A and B respectively. A flow of purple color in the middle of the layer depicts shared keywords within a theme or topic. Each stream in the graph is characterized by a color transition from blue to red from the top to the lower border. The extreme blue and red at the borders become the two poles represent each text corpora, the name of this visualization technique, Polarized Streamgraph, is named after the polarized effect of layers/streams in the graph. If a stream is dominated by either blue or red color, one can interpret the result as either one corpus is leading a theme over the other.

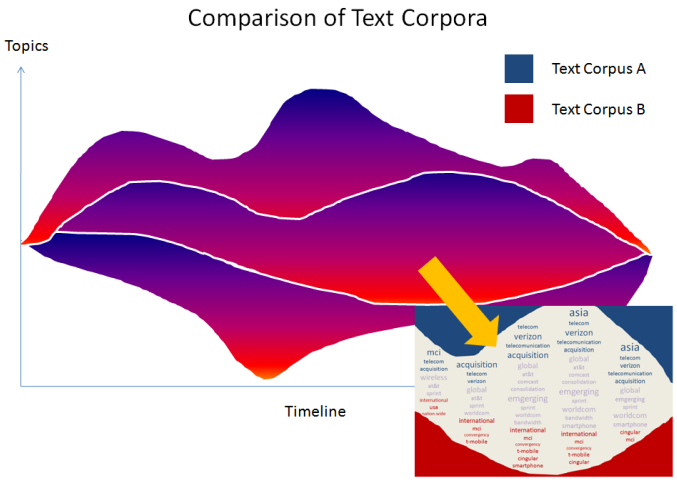


Figure 1 – Polarized Streamgraph for Comparison of Text Corpora

For constructing the stream graph, we use one of the methods by Byron & Wattenberg [2] to compute the base line for the streamgraph with the following equation to balance the computational cost and the aesthetics effect of the streamgraph.

$$g_0 = -\frac{1}{n+1}\sum_{i=0}^{n}\sum_{j=1}^{i} f_j = -\frac{1}{n+1}\sum_{i=1}^{n}(n-i+1)f_i$$

where g_0 is the baseline value, and f_i is the i-th time series and n is the total number of series. In our approach, we use distinct rectangles instead of an interpolation to represent the keyword information. We refer to these rectangles as "keyboxes". An interpolation of the color transition generates better aesthetic visual effect whereas our approach provides more accurate representation of the keyword occurrence from the compared text corpora. The height of the keyboxes encodes a keyword's frequency of occurrence. The color of the keyboxes represents which text corpus dominates for the keyword, i.e. the pure red depicts the keywords are unique to corpus B, the pure blue are unique keywords from corpus A. The common or shared keywords are illustrated by the reddish purple or bluish purple keyboxes defined by the following equation:

$$C_r = f_1/(f_1 + f_2)\times 255$$
$$C_b = 255 - C_r$$

where C_r and C_b are the red and blue components of the RGB color model respectively. Each keybox represents one keyword in a particular time interval, f_1 and f_2 are the frequency of the keyword for corpus A and corpus B respectively, the number 255 represents the maximum intensity of the color components.

Keywords in a topic in a time interval are sorted by the ratio of the frequency counts in the two corpora. We combine all these keyboxes to form a stream and the streams from various topics complete the streamgraph as shown in the following magnified diagram.

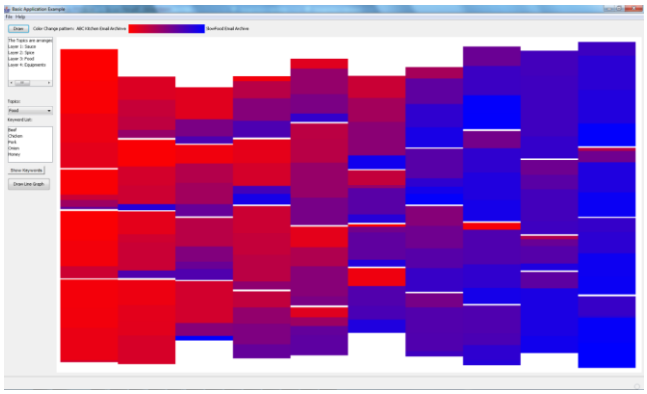


Figure 2 – Magnified streams of Polarized Streamgraph

To further detail the comparison, we implemented a coordinated window to support the drilldown of any topic to the keyword-level. In the coordinated window, we extend the parallel tag cloud idea to build the columns of words corresponding to their occurrence related to the timeline. Keywords in blue and red colors are unique to corpus A and B. The keywords highlighted with purple color are shared among the two corpora. The size of the keywords encodes the frequency of their occurrences in the text as illustrated in the following diagram.

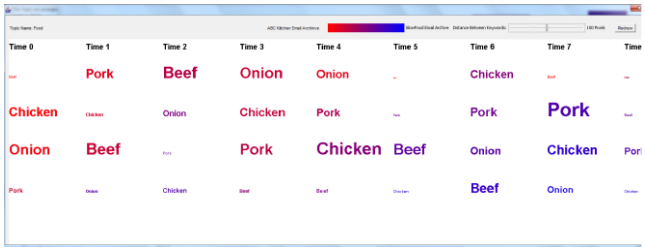


Figure 3 – Keyword-level View

5. FUTURE WORK & CONCLUSION

After a survey on email visualization techniques, we find there is no previous attempt to visualize discussion trend or change of themes over time at the institutional level. It is both interesting and valuable to explore hot topics or themes over time in a company. The adoption of semantic analysis technique is inevitable as the visualization of only the information embedded in the message headers provides very little insight into the knowledge hidden in any email archives. Unknown knowledge hidden in email archives is important to research in corporate culture, organizational behavior, competitive analysis and sometimes crime investigations. Trampoline's Enron Explorer provides a great platform to visualize the overall communication pattern and themes within Enron. However, it lacks the encoding of temporal information in its visualization. It does not support the visualization of change of themes over time or change in communication pattern over time. We introduce an extension of streamgraph for comparing text corpora. Our approach allows a bi-level visual examination of any two text corpora at both the theme/topic level and at the keyword-level with the use of a coordinated window. The polarized streamgraph provides a holistic view of similarities and differences between text corpora over time. This visual can illustrate clearly the onset and die off of hot topics. One can use our method to compare an institutional email archive against market news, or to contrast discussion trends between two institutions. One major limitation of our current design is no intuitive extension of the polarized streamgraph to compare more than two text corpora at the same time. An interesting possible improvement is the development of an incremental algorithm to support the modeling of topics in near to real-time to support analysis of data streams such as status updates on facebook or twitter.

6. REFERENCES

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. J.Mach.Learn.Res., 3, 993-1022.

[2] Byron, L., & Wattenberg, M. (2008). Stacked graphs - geometry & aesthetics. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1245-1252.

[3] Clark, J. (2008). Obama McCain Convention Speech Comparison. http://www.neoformix.com

[4] Collins, C., Viégas, F.B., Wattenberg, M. (2009). Parallel Tag Clouds to explore and analyze faceted text corpora. Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on , vol., no., pp.91-98, 12-13 Oct.

[5] Fisher, D., Brush, A. J., Gleave, E., & Smith, M. A. (2006). Revisiting whittaker & sidner's email overload ten years later. CSCW '06: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, Banff, Alberta, Canada. 309-312.

[6] Fisher, D., Hogan, B., Brush, A. J., Smith, M. A., & Jacobs, A. (2006). Using social sorting to enhance email management. Proc. Human-Computer Interaction Consortium (HCIC).

[7] Frau, S., Roberts, J. C., & Boukhelifa, N. (2005). Dynamic coordinated email visualization. WSCG05 - 13th International Conference on Computer Graphics, Visualization and Computer Vision'2005, 187-193.

[8] Gemmell, J., Bell, G., & Lueder, R. (2006). MyLifeBits: A personal database for everything. Commun.ACM, 49(1), 88-95.

[9] Liu, S., Zhou, M. X., Pan, S., Qian, W., Cai, W., & Lian, X. (2009). Interactive, topic-based visual text summarization and analysis. CIKM '09: Proceeding of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China. 543-552.

[10] Paley, W. B. (2002). TextArc: Showing Word Frequency and Distribution in Text. Interactive poster presented at InfoVis '02.

[11] Perer, A., Shneiderman, B., & Oard, D. W. (2006). Using rhythms of relationships to understand e-mail archives. J.Am.Soc.Inf.Sci.Technol., 57(14), 1936-1948.

[12] Perer, A., & Smith, M. A. (2006). Contrasting portraits of email practices: Visual approaches to reflection and analysis. AVI '06: Proceedings of the Working Conference on Advanced Visual Interfaces, Venezia, Italy. 389-395.

[13] Rohall, S. L., Gruen, D., Moody, P., Wattenberg, M., Stern, M., Kerr, B., et al. (2004). ReMail: A reinvented email prototype. CHI '04: CHI '04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria. 791-792.

[14] Salton, G., & McGill, M., editors. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

[15] Thorp, J. (2009). Two Sides of the Same Story: Laskas & Gladwell on CTE & the NFL. http://blog.blprnt.com

[16] Trampoline Systems. 2006. Enron Explorer. http://www.enronexplorer.com

[17] Viégas, F. B., Boyd, D., Nguyen, D. H., Potter, J., & Donath, J. (2004). Digital artifacts for remembering and storytelling: PostHistory and social network fragments. HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4, 40109.1.

[18] Viégas, F. B., Golder, S., & Donath, J. (2006). Visualizing email content: Portraying relationships from conversational histories. CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, Québec, Canada. 979-988.

[19] Whittaker, S., & Sidner, C. (1996). Email overload: Exploring personal information management of email. CHI '96: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, British Columbia, Canada. 276-283.

Visual Content Correlation Analysis

Furu Wei[1], Lei Shi[1], Li Tan[1], Xiaohua Sun[1], Xiaoxiao Lian[1], Shixia Liu[1], Michelle X Zhou[2]
IBM Research
[1]{weifuru, shllsh, tltan, sunxiaoh, liusx, xxlian }@cn.ibm.com, [2]mzhou@us.ibm.com

ABSTRACT
Correlating content from multiple data fields is one of the key challenges in text mining. In this paper, we propose a visual analytics approach that leverages both content correlation analysis and interactive visualization technologies in analyzing and understanding content correlations. We have applied our work to analyzing NHAMCS data (National Hospital Ambulatory Medical Care Survey), which helps reveal healthcare-related data patterns through the correlations between unstructured data fields (e.g., cause of injury and diagnosis) and between structured and unstructured fields (e.g., gender and cause of injury).

Author Keywords
Content correlation, visualization, visual text analytics.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Modeling documents in a large text corpus into a set of topics is an effective method for information digesting. However, there are situations in which this method by itself may not be enough. In many real-life applications, we may deal with a collection of documents that are from multiple sources (e.g., news, forum) or each of which consists of multiple fields describing the same data entry from different perspectives. In both cases, a further analysis of the correlation among the contents from different sources/fields can help users gain more insights from the corpus under analysis. For example, the record of an emergency room visit has the free-text fields about *the cause of the injury*, *the reason of visit*, and *the diagnosis*. It also has structured information, such as the gender or the age of the patient. While analyzing such a data set, a health insurance analyst may want to find out the relationship between *the reason of visit* and the patient groups divided by the age values (one unstructured and one structured field), doctors may want to detect and accumulate more knowledge about the relationship between a symptom and its potential causes (two unstructured fields).

In this paper, we propose a visual analytics approach to supporting the above mentioned content correlation analysis tasks. Our previous work TIARA [3] is used as a starting framework. TIARA provides a visual summary of a text corpus in the form of an enhanced stacked graph. To support content correlation analysis, we further augment TIARA from two aspects. From the analysis side, we added techniques supporting multi-field content correlation analysis. From the visualization side, we developed a more robust sweep-line based keyword layout method and introduced a customized fisheye distortion to allocate more space for visually illustrating the content correlation of topics under focus.

Our major contributions are:

1) We design a time-based, topic-oriented visualization to visually illustrate the correlation analysis results learned from multiple fields/sources of a text corpus.
2) We provide rich context and interactions to help users understand the analytic results from multiple perspectives in order to compensate the deficiencies of current content correlation analysis technologies.

RELATED WORK
Researchers have developed correlated topic models with different foci. Salomatin et al. [4] found the need for analyzing data collections whose entries consist of multiple fields with different but interrelated contents. They also extended CTM [2] to model the correlated topics from multi-fields of a heterogeneous data corpus. However, their research focuses on the modeling of correlated topics instead of analyzing the correlation of content learned from multiple fields in each topic. Similarly, the research by Wang et al. [5] also focuses on mining the correlated topic patterns rather than revealing the correlation of topic content from different sources.

Compared with existing works, we proposed an effective visual analytic method to visually analyze the content correlation learned from multiple fields/sources. Our work seamlessly integrates content correlation analysis techniques with interactive visualization to support an iterative and progressive text analysis.

VISUAL CONTENT CORRELATION ANALYSIS
The major feature of our work is that it tightly integrates text analytics and interactive visualization to help users better consume complex analytic results. It first visualizes the analytic results to help users understand them. Then upon users' interaction with the visualization, it provides more analysis support and displays the results from further analysis. In this section, we'll describe our work from both aspects.

Topic based Content Correlation Analysis
To help users analyze the correlation among contents from multiple fields/sources, we develop a topic based content correlation analysis method. It involves two main tasks: (1)

modeling the text collection into a set of topics with time sensitive keywords illustrating the content evolution over time; (2) analyzing the content correlation among different fields/sources.

Topic annotation and keyword based summarization

In one application of our system to a multi-field data corpus (National Hospital Ambulatory Medical Care Survey (NHAMCS) data from U.S. CDC), we first trained a unified topic model by combining the multiple text fields in each document, and then select topic keywords separately for each field. Specifically, LDA [1] is used to annotate the topics in a document collection. Given a document collection $D = \{d_1, d_2, \cdots d_N\}$ with N as the document number, each document d_i is assigned with a distribution over the topics $T = \{t_1, t_2, \cdots t_K\}$ with K as the pre-defined topic number. Meanwhile, each topic t_j is also assigned a distribution over the word vocabulary $W = \{w_1, w_2, \cdots w_M\}$ of the whole text collection, where M is the size of the vocabulary. We select the top 50 keywords to summarize each topic in terms of the probability of the keyword belonging to the topic. Moreover, to detect and reveal the topic evolution, the topics are further split into several parts along the time dimension, and then the top 15 most frequent terms are selected as the time-sensitive keywords for each time range.

Multi-field content correlation analysis

Let $f_i = f_{i1}, f_{i2}, \cdots f_{iO}$ denote the text fields in document d_i. After the LDA based topic annotation by combining f_i into a whole text document, it requires a strategy to assign a topic label for each word in the multiple text fields. We already generate a topic label for each word in the word sequence from D. Since the boundary for each field in one document is available, the topic labels for words coming from different fields can be induced directly. For the task of content correlation analysis, we extract the top 15 frequent words as time-sensitive summarization for each field in each time range, through which the user can observe and analyze the content correlation among different fields of the text collection.

Topic and keyword re-ranking

The topics derived by LDA are randomly ordered, thus it is often useful to order the topics so that the most important ones are shown first. Specifically, we discount topics that are common to many documents in a document collection. Mathematically, it can be formulated as follows. Let $\theta_i = \{\theta_{i1}, \theta_{i2}, \cdots \theta_{iN}\}$ be the distribution of the topic t_i in all the documents, then the rank of t_i is computed by: $\hat{\theta}_i = mean(\theta_i) \times dev(\theta_i)$, where $mean(\theta_i) = \sum_i \theta_i / N$, and $dev(\theta_i) = \sqrt{\sum_i (\theta_i - mean(\theta_i))^2 / (N-1)}$.

Furthermore, to help user better understand the topic analysis results, we re-rank topic keywords to enhance the topic definitions. Inspired by the classical TFIDF keyword weighting scheme in information retrieval, the rank of a keyword w in topic t_i is determined by, $\eta_w = \phi(w, t_i) \times \log(K / \varphi(w, T))$, where $\phi(w, t_i)$ denotes the frequency of w in t_i, and $\varphi(w, T)$ denotes the number of topics which contain w.

Content Correlation Visualization

We add to TIARA the following interactive visualization features to support the content correlation analysis:

1) Providing more space for the topic under inspection,

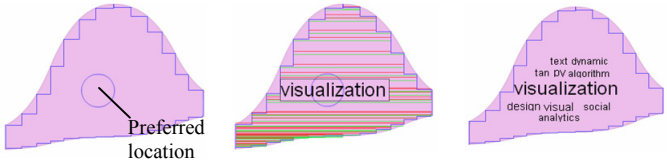
2) For revealing correlation among multiple unstructured data fields/sources, dividing the topic stripe under focus into multiple sub-stripes to display keywords extracted from each field,

3) For revealing correlation between structured and unstructured fields, at the overview level, brushing the topic keywords from the unstructured field(s) using the categorical values from the structured field; at the detailed level, dividing the focused topic stripe into sub-stripes according to the structured field and populating the keywords from the structured field(s) into the sub-stripe for each category correspondingly.

In the above designs, time-sensitive keywords summarized from each field for the topics play an important role in displaying the correlation of contents from different fields at specific time point on a specific topic. A robust method for packing the time-sensitive keyword clouds and an intuitive interaction mechanism for allocating more keyword display space for the topic under inspection are thus critical for the visual content correlation analysis. For these reasons we choose to focus on the following two aspects research-wise.

Keyword cloud layout

We propose a sweep-line based approach to placing the keywords (see Fig.1). The approach works in greedy manner; and it lays out each keyword one by one.



(a) generate the inner boundary (b) sweep-line based search (c) final result

Fig 1. Keyword cloud layout

In order to simplify the region constraints and facilitate the sweep-line based search process, we approximate the boundary curves of each topic layer by a rectilinear polygon (a polygon all of whose edges meet at right angles). The rectilinear polygon is constructed by points sampled from the curve boundary. A preferred location is then assigned to the keyword cloud in a given region. In our implementation, it is the center of given region. Fig 1(a) shows the rectilinear polygon and the preferred location for the given region.

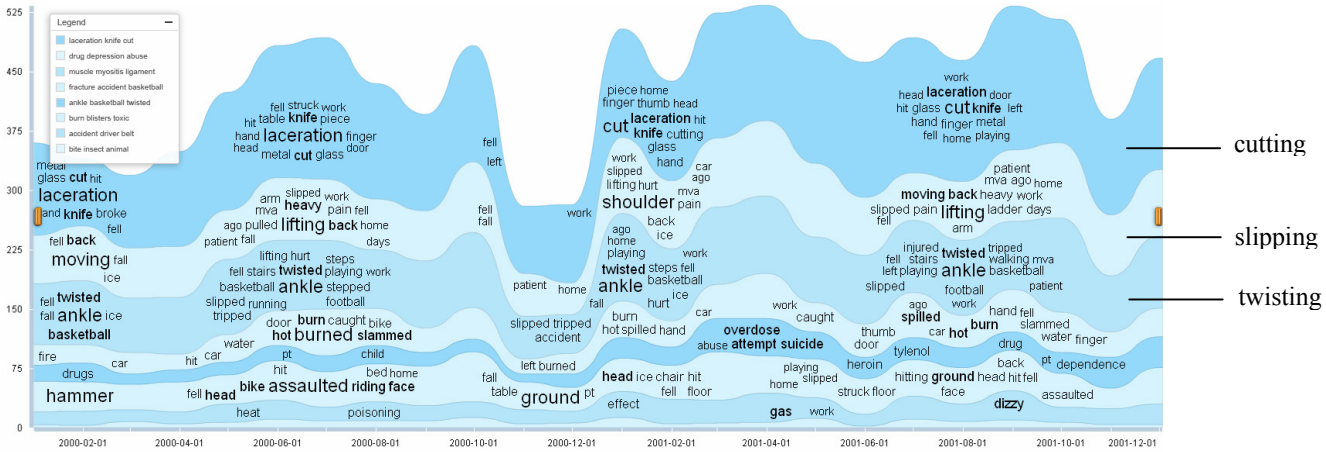


Fig.2: Visual summary of single text field (cause of injury) for healthcare data

To generate a compact word cloud, our layout method places each keyword as close as possible to the preferred location in the available space of the rectilinear polygon. This is achieved by the following steps. First, for each keyword, we search the best location by the sweep line technique. Here, only the vertices of the rectilinear polygon are used in the sweep line search process. Then, the rectilinear polygon is dynamically updated according the bounding box of the new placed keyword. Fig 1(b) illustrates the layout of the first keyword, and Fig 1(c) shows the final layout.

Fisheye distortion

To enable the user to observe more details of topic content correlation, we introduce customized fisheye distortion [6] to enlarge the topic of interest. Compared with the traditional fisheye distortion, our approach varies and extends in two aspects: First, different from the typical technique of the fisheye distortion, we do not enlarge the focused stripe by the pre-defined fisheye parameter uniformly. Instead, we proactively compute the required area to place our keyword cloud and adaptively set the fisheye parameter to fit the available space. Hence, we can have enough space to place more keywords for the thin stripe area. Second, we introduce a post-adjusting method to maintain the relative height of the topic stripe along the time dimension, so that the visual momentum is not affected by the fisheye distortion.

EXAMPLE SCENARIO

In this section, we use a concrete example in healthcare domain to show how our visual analytics system can help users in their content correlation analysis. The NHAMCS data we based on includes both free-text and categorical information of sample cases in ambulatory care services provided by hospital emergency and outpatient departments in U.S. Our system analyzes about 20,000 records during the years of 2000 and 2001.

Consider Alice, a physician in her first year of residence, is studying the existing emergency room records in order to gain more knowledge about the interrelated patterns among different facets of emergency room visits. She starts by asking our system to give her a visual summary of all the records under inspection. Responding to her request, a visualization with stripes representing the temporal evolvement of top eight topics is created. She then selects the "Cause of Injury" item, an unstructured text field included in the records, to see the cause of injury information for each topic. As illustrated in Fig. 2, she can see that the main cause of injury for the top three record topics are "cutting", "slipping" and "twisting" correspondingly.

Now Alice is interested in knowing whether there is any difference between the causes of injury for men and women for each topic. She then selects the "Patient Gender" filed, a structured data field from the records. The keywords for each topic are then brushed based on the gender values, where dark green indicates the keywords more frequently found in the male patient records, such as "cut" and "knife"; and brown represents those appears more in the female patient records, such as "burned" and "ground". In order to see more details about the topic which has many records with twist-related cause of injury, Alice clicks on the corresponding topic stripe (the third in top-down order) to drill-in. As displayed in Fig. 3, the selected stripe is enlarged using our customized fisheye distortion technique. The stripe is also further divided into two sub-stripes for displaying keywords from each gender category. In this view, some interesting patterns are revealed. The injury causes of male patients tend to be due to competition-related activities, such as playing football, basketball, and wrestling; whereas for female patients, the injuries often happen during everyday life activities, such as walking and running.

After exploring the correlation between the cause of injury and the gender fields, Alice switches to looking at the diagnosis other doctors gave to those visits with the cause of injuries she just studied. This time she selects both "Cause of Injury" and "Diagnosis", two unstructured fields. Again, she drills down into that topic which has many records with twist-related causes. From the visualization returned (Fig. 4) Alice can clearly see that injuries caused by twist-related reasons are mostly diagnosed as strains/sprains and in most cases happen to the body parts like ankle, wrist or knee.

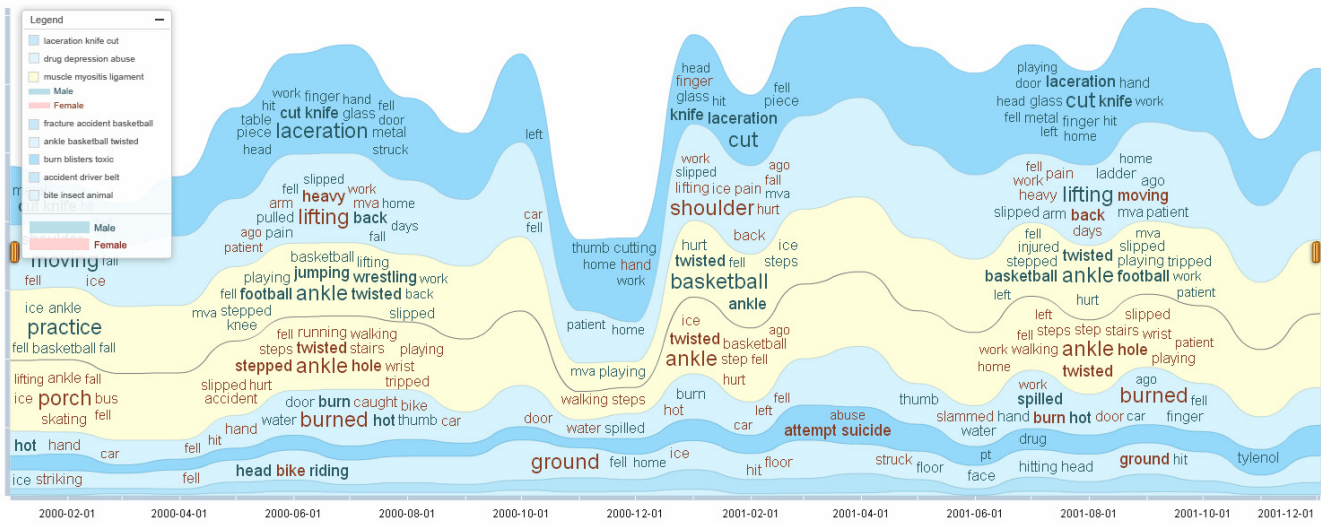


Fig. 3: Content correlation view of one structured field (gender) and one unstructured field ("cause of injury")

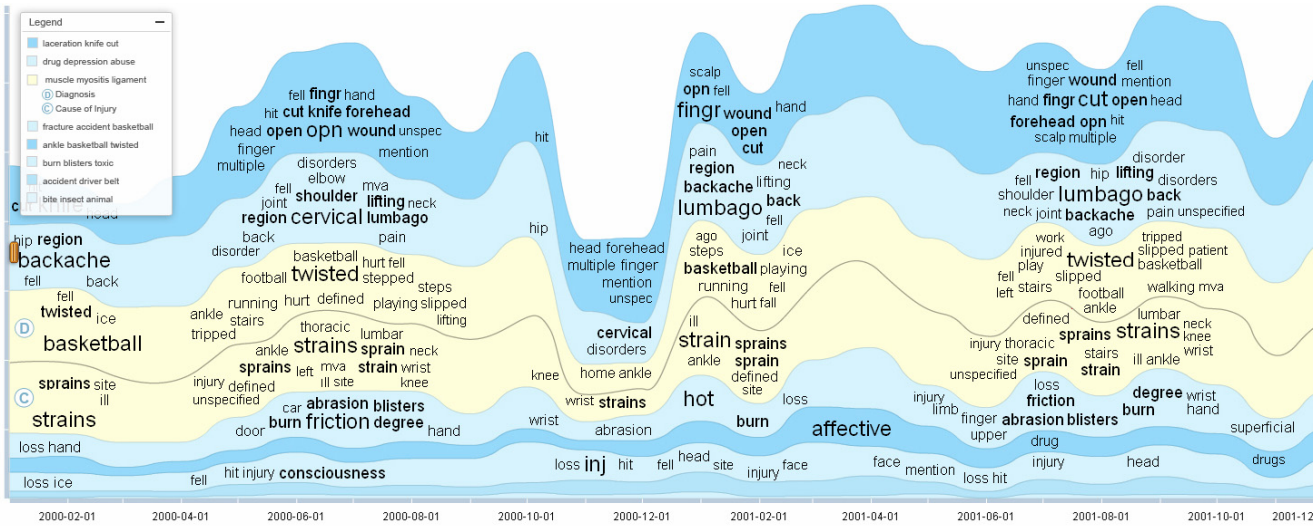


Fig. 4: Content correlation view of two text fields ("cause of injury" and "diagnosis")

CONCLUSIONS

In this paper, we have presented our preliminary work on content correlation analysis through a visual analytics approach. Specifically, we developed a system that tightly integrates text analytics with interactive visualization and developed supporting technologies for both aspects. Our approach has two main advantages: 1) It makes the correlation analysis result consumable by leveraging the interactive visualization techniques. 1) It leverages human visual perceptual capability to compensate for the deficiencies of the correlation analytics techniques.

REFERENCES

1. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 (2003): 993-1022.
2. David M. Blei and John D. Lafferty. A Correlated Topic Model of Science. *The Annals of Applied Statistics* Vol. 1 (2003), No. 1 (2003), 17–35.
3. Shixia Liu, Michelle X Zhou, Shimei Pan, Weihong Qian, Weijia Cai, Xiaoxiao Lian. Interactive, Topic-based Visual Text Summarization and Analysis. In *Proc. CIKM 2009*, 543-552.
4. Konstantin Salomatin, Yiming Yang, and Abhimanyu Lad. Multi-field Correlated Topic Modeling. In *Proc. SDM 2009*, 628-637.
5. Xuanhui Wang, ChengXiang Zhai, Xiao Hu, Richard Sproat. Mining Correlated Bursty Topic Patterns from Coordinated Text Streams. In *Proc. KDD 2007*, 784-793.
6. M. Sarkar and M. Brown. Graphical fisheye view. Comm. ACM, 37(12):73–83, 1994.

An Ontology-based Interface for Improving Information Exploration

Wilson Wong
Centre for Software Practice
University of Western Australia
wilson.wong@csp.uwa.edu.au

Wei Liu and Mohammed Bennamoun
School of Computer Science and Software Engineering
University of Western Australia
{wei,bennamou}@csse.uwa.edu.au

ABSTRACT

The explosive growth of the Web necessitates the use of approaches such as search engines, social indexing and visualisation systems to assist in exploring online information. Current systems based on keyword search and sequential listing are becoming inadequate due to their inability to make sense of unstructured information. This calls for more work into the automatic structuring of text to produce accurate and machine-friendly metadata. This paper presents a working ontology-based interface for exploring large amounts of news articles across different domains based on the seamless and automatic discovery of document abstractions.

ACM Classification Keywords

H.5.0 Information Interfaces and Presentation: General; I.2.7 Artificial Intelligence: Natural Language Processing—*text analysis*; H.3.1 Information Storage and Retrieval: Content Analysis and Indexing—*abstracting methods*

1. INTRODUCTION

The explosion of textual information on the Web (i.e. *information explosion*) places great stress on our cognitive abilities. Research has shown that people have short attention span on the Web [5], and are slow at reading off the screen [4]. For these reasons, users are finding it increasingly difficult to explore the excessive amount of information available, an effect known as *information overload*. This problem becomes more alarming when we consider the fact that more than 90% of the data in the world appear in unstructured forms [3].

As a result, providing users with more effective ways of exploring mountains of information on the Web has become a challenge for Web developers and researchers alike. Many existing efforts towards this end remain within the realm of theoretical contributions with shortcomings to be discussed in Section 2. We have developed a two-part solution to this challenge. This paper presents a working ontology-based interface[1] (i.e. the first part of the solution) for improving information exploration using two types of abstractions, namely, term clouds and lightweight ontologies. These abstractions are automatically generated using our ontology learning system (i.e. the second part of the solution) as shown in Figure 1. More details about this system are available in [9]. The four main distinguishing points of this system are (1) the process of discovering abstractions is fully automatic, (2) the system depends solely on dynamic Web data, (3) the system works across different domains (our current focus is technology, economics and medicine), and (4) the individual techniques in the system have been evaluated to produce highly accurate and complete abstractions. Using three scenarios, Section 3 discusses how our interface enables users to quickly identify the **overall ideas** or **specific information** in **individual documents** or large **groups of documents**.

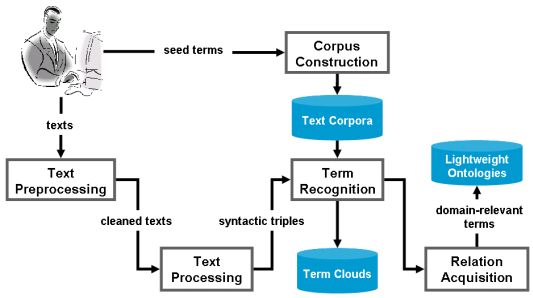


Figure 1. An overview of our ontology learning system for discovering term clouds and lightweight ontologies from documents across different domains.

2. RELATED WORK

Various efforts have been undertaken to improve the exploration of textual information. These efforts range from being highly dependent on human involvement to automatic systems. **First**, there have been several well-financed efforts on advanced graphical user interfaces for organising and displaying information (e.g. news, search results). Despite the fancy interfaces, most of these attempts have limited success primarily due to users' adherence to text as the predominant medium of communication on the Web [7]. **Second**, clustering techniques such as Scatter/Gather [2] have been proposed to organise documents into groups to aid browsing.

[1] http://explorer.csse.uwa.edu.au/research

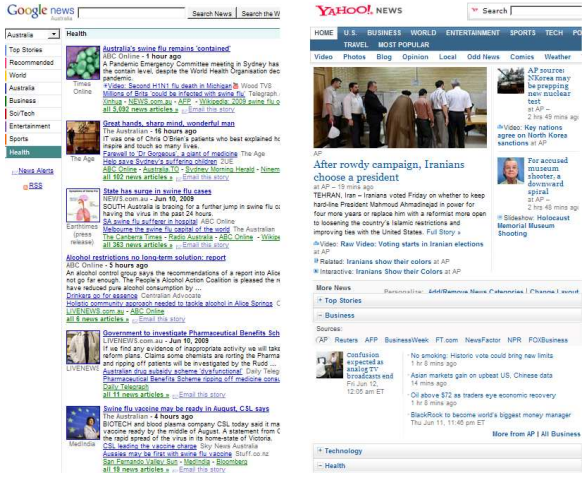


Figure 2. The screenshot of the aggregated news services provided by Google (the left portion of the figure) and Yahoo (the right portion of the figure) on 11 June 2009.

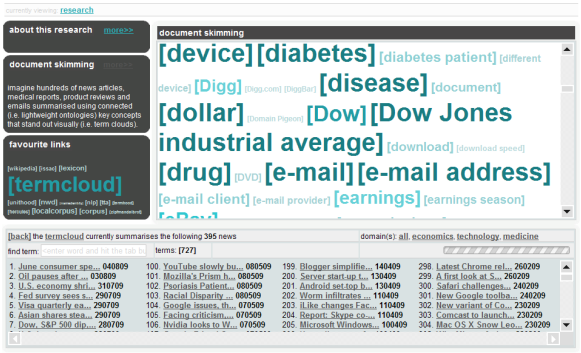


Figure 3. The cross-domain term cloud summarising the main concepts occurring in all the 395 articles listed in the news browser. This cloud currently contains terms in the technology, medicine and economics domains.

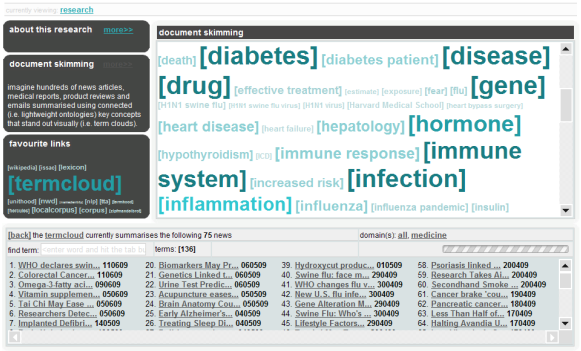


Figure 4. The single-domain term cloud for the domain of medicine. This cloud summarises all the main concepts occurring in the 75 articles listed below in the news browser. Users can arrive at this single-domain cloud from the cross-domain cloud in Figure 3 by clicking on the `[domain(s)]` option in the latter.

However, processing time and scalability remain an issue. **Third**, collaborative tagging, which is the labelling of Web resources (e.g. news articles, photos) by end users using Web 2.0 services, is considerably more successful. However, the usefulness of folksonomies are often questioned [8] due to the biased and non-standardised creation of tags. **Fourth**, the more ambitious Semantic Web vision aims to enable computers to automatically find information and services on the Web on behalf of users using machine-understandable metadata. However, practical concerns such as the reliance on humans to create the metadata [6] have become a major bottleneck. **Lastly**, such concern for manual curation of metadata has given rise to automatic techniques for extracting structured data from text. For instance, `Yahoo! Term Extraction` is an automatic service for extracting terms from cross-domain texts. [1] proposed the ScentIndex/ScentHighlights techniques based on co-occurrence analysis and lexical matching to retrieve and highlight information.

3. ONTOLOGY-BASED INFORMATION EXPLORATION

As summarised in Section 2, existing approaches are typically too superficial based on lexical-level and statistical techniques (incomplete, inaccurate output), or rely too much on static background knowledge and manually-created rules (non-portable, non-scalable output). More importantly, these techniques are mostly directed towards information access at the collection level to complement more focused approaches.

In this section, we discuss three use cases to demonstrate how the automatically discovered abstractions (i.e. term clouds and lightweight ontologies) by our system in Figure 1 can be employed to assist users to explore texts more effectively, at both the collection and the document level. We take the two aggregated news services by Google and Yahoo shown in Figure 2 as examples to demonstrate the current lack of attention on information exploration. The left portion of the figure shows the Google News interface, while the right portion shows the Yahoo News interface. Both interfaces are focused on the health news category on 11 June 2009. The actual listings are considerably longer. A quick look at both interfaces would immediately reveal the tremendous time and cognitive effort that users have to invest in order to arrive at a summary of the texts or to find a particular piece of information. To address this, we introduce a new exploration interface based on term clouds and lightweight ontologies to assist users in quickly identifying the **overall ideas** or **specific information** in **individual documents** or **groups of documents**. In particular, the following three cases are examined:

(1) Can the users quickly guess (in 3 seconds or so) from the listing alone what are the **main topics of interest across all articles** for that day?

(2) Is there a better way to present the **gist of individual news articles** to the users other than the conventional, ineffective use of short text snippets as summaries?

(3) Are there other options besides the typical `[find]` feature for users to quickly **pinpoint a particular concept** in **an article** or **a group of articles**?

While news articles may be the focus of the current explo-

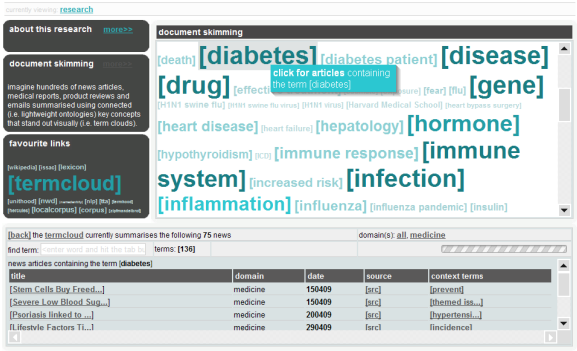


Figure 5. The single-domain term cloud for the medicine domain. Users can view a list of articles describing a particular topic by clicking on the corresponding term in the single-domain cloud.

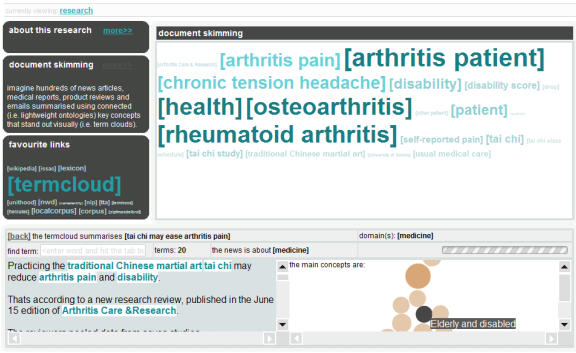


Figure 6. The document term cloud for the article *"Tai Chi may ease arthritis pain"*. The document term cloud and information from the lightweight ontology offer an innovative and intuitive way of presenting the gist of individual news articles.

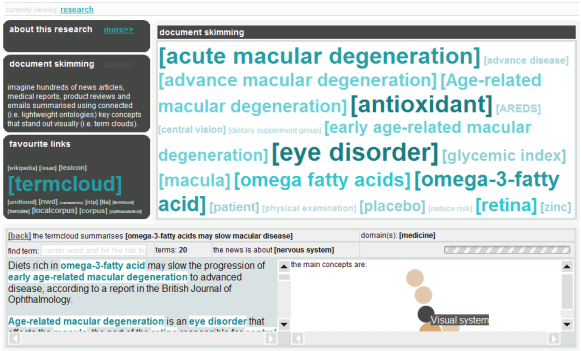


Figure 7. The document term cloud for the article *"Omega-3-fatty acids may slow macular disease"*. Based on the term size in the clouds, one can arrive at the conclusion that the news featured in Figure 7 carries more domain-relevant (i.e. medical related) content than the news in Figure 6.

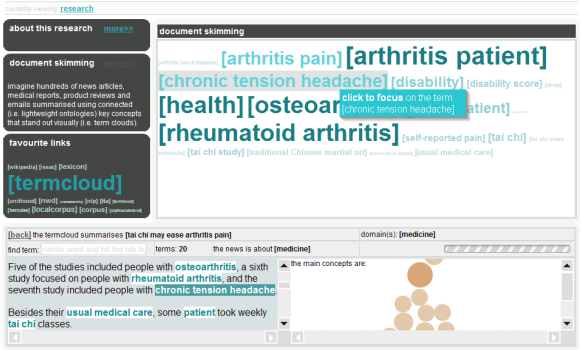


Figure 8. The document term cloud for the news *"Tai Chi may ease arthritis pain"*. Users can focus on a particular concept in the annotated news by clicking on the corresponding term in the document cloud.

ration interface, other text documents including product reviews, clinical notes, emails and search results can equally benefit from our two-part solution. Figure 3 shows the main interface for exploring a list of news articles across different domains. The white canvas on the top right corner containing words of different colours and sizes is the *cross-domain term cloud*. This term cloud summarises the key concepts in all news articles across all domains listed in the *news browser* panel below. For instance, Figure 3 shows that there are 395 articles across three domains (i.e. technology, medicine and economics) listed in the news browser with a total of 727 terms in the term cloud. The solutions to the above three use cases using our ontology-based interface are as follows.

Case 1: Figure 4 shows the *single-domain term cloud* for summarising the key concepts in the medicine domain. This term cloud is obtained by simply selecting the `medicine` option in the `[domain(s)]` field. There are 75 articles in the news browser with a total of 136 terms in the cloud. Looking at this single-domain term cloud, one would immediately be able to conclude that some of the news articles are concerned with *"diabetes"*, *"drug"*, *"gene"*, *"hormone"*, *"heart disease"*, *"H1N1 swine flu"* and so on. One can also say that *"diabetes"* was discussed more intensely in these articles than other topics such as *"heart disease"*. The users are able to **grasp the gist** of **large groups of articles** in a **matter of seconds** without any **complex cognitive effort**. Can the same be accomplished through the typical news listing and summaries shown in Figure 2? The use of the cross-domain or single-domain term clouds for summarising the main topics across multiple documents addresses the first problem.

Case 2: If the users are interested in drilling down on a particular topic, they can do so by simply clicking on the terms in the cloud. A list of news articles describing the selected topic is provided in the news browser panel as shown in Figure 5. The context in which the selected topic exists is also provided. For instance, Figure 5 shows that the *"diabetes"* topic is mentioned in the context of *"hypertension"* in the news *"Psoriasis linked to..."*. Clicking on the `[back]` option brings the users back to the complete listing of articles in the medicine domain as in Figure 4. The users can also preview the gist of a news article by simply clicking on the title in the news browser panel. Figure 6 and 7 show the *document term clouds* for the news *"Tai Chi may ease arthritis pain"* and *"Omega-3-fatty acids may slow macular disease"*. These document term clouds summarise the content of the news articles and present the key terms in a visually appealing manner to enhance the interpretabil-

ity and retention of information. The interfaces in Figure 6 and 7 also provide information derived from the corresponding lightweight ontologies. For instance, the root concept in the ontology is shown in the [this news is about] field. In the news *"Omega-3-fatty acids may slow macular disease"*, the root concept is *"nervous system"*. The parent concepts of the key terms in the ontology are presented as part of the field [the main concepts are]. In addition, based on the term size in the clouds, one can arrive at the conclusion that the news featured in Figure 7 carries more domain-relevant content (i.e. more medically-related) than the news in Figure 6. Can the users arrive at such **comprehensive and abstract information** regarding a document with **minimal time and cognitive effort** using the conventional interfaces shown in Figure 2? The use of document term cloud and lightweight ontologies for presenting the gist of individual news articles addresses the second problem.

Case 3: The use of the following features [click for articles], [find term], [context terms] and [click to focus] helps users to locate a particular concept at different level of granularities. At the document collection level, users can locate articles containing a particular term using the [click for articles], the [find term] or the [context terms] features. The [click for articles] feature allows users to view a list of articles (using the news browser) related to a particular topic in the cross-domain or the single-domain term cloud. The [find term] feature can be used anytime to refine and reduce the size of the cross-domain or the single-domain term cloud. Context terms are provided together with the listing of articles in the news browser when users select the [click for articles] feature. Clicking on any terms under the column [context terms], as shown in Figure 5, will list all articles containing the selected term. At the individual document level, news articles are annotated with the key terms that occurred in the document clouds to assist scanning activities. Users can employ the [click to focus] feature to pinpoint the occurrence of a particular concept in an article by clicking on the corresponding term in the document cloud. Figure 8 shows how a user clicked on *"chronic tension headache"* in the document term cloud which triggered the auto-scrolling and highlighting of that term in the annotated news. Can the users **pinpoint a particular topic** that occurred in a **large document collection** or a **single lengthy document** with **minimal time and cognitive effort** using the conventional interfaces shown in Figure 2? The various features provided by this system allow users to quickly pinpoint a particular concept, either in an article or a group of articles, to address the last problem.

4. CONCLUSION

Information overload has become an inevitable part of our daily lives. The inability of existing approaches to assist users in exploring mountains of information aggravates the situation. We employed our ontology learning system to automatically discover two types of document abstractions, namely, term clouds and light-weight ontologies. These abstractions are used by our working ontology-based interface to assist users in exploring information more effectively, at both the collection and the document level. The benefits of using automatically generated term clouds and lightweight ontologies for exploring information were highlighted using three use cases.

Qualitatively, the three use cases demonstrated that conventional news listing interfaces, unlike our ontology-based information exploration interface, are unable to satisfy the following three common scenarios: (1) to grasp the gist of large groups of articles in a matter of seconds without any complex cognitive effort, (2) to arrive at a comprehensive and abstract overview of a document with minimal time and cognitive effort, and (3) to pinpoint a particular topic that occurred in a large document collection or a single lengthy document with minimal time and cognitive effort. Since we are at the early stages of the development process, this paper reports only the preliminary usability inspection using several use cases. The effectiveness of the ontology-based interface will be further assessed using quantitative means.

REFERENCES

1. E. Chi, L. Hong, J. Heiser, S. Card, and M. Gumbrecht. Scentindex and scenthighlights: Productive reading techniques for conceptually reorganizing subject indexes and highlighting passages. *Information Visualization*, 6(1):32–47, 2007.
2. D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 1992.
3. J. Gantz. The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. Technical Report White paper, International Data Corporation, 2008.
4. J. Gould, L. Alfaro, R. Finn, B. Haupt, A. Minuto, and J. Salaun. Why reading was slower from crt displays than from paper. *ACM SIGCHI Bulletin*, 17(SI):7–11, 1986.
5. S. Krug. Dont make me think! a common sense approach to web usability. New Riders, Indianapolis, USA, 2000.
6. C. Marshall and F. Shipman. Which semantic web? In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, Nottingham, UK, 2003.
7. P. Morville. Ambient findability. OReilly, California, USA, 2005.
8. C. Shirky. Folksonomies + controlled vocabularies. http://many.corante.com/archives/2005/01/07/folksonomies_contro 2 September 2009, 2005.
9. W. Wong. *Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge*. PhD thesis, University of Western Australia, 2009.

Information Visualization for Corpus Linguistics: Towards Interactive Tools

Harri Siirtola, Kari-Jouko Räihä
TAUCHI
Department of Computer Sciences
University of Tampere
{harri.siirtola,kari-jouko.raiha}@cs.uta.fi

Tanja Säily, Terttu Nevalainen
VARIENG
Department of English
University of Helsinki
{tanja.saily,terttu.nevalainen}@helsinki.fi

ABSTRACT
In this paper linguists and researchers of visual data analysis outline the requirements and benefits of an information visualization approach for corpus linguistics. Over the years, the information visualization community has come up with a number of methods to visualize text, but the majority of these techniques do not serve the needs of the linguistic community. This is evident in the over-simplification of the linguistic problems and generally caused by a poor understanding of the domain. We started a joint research effort with linguists, data miners, and information visualizers to design and produce better data analysis tools for corpus linguistics. This work is still in its early stages, but we have a shared vision of what needs to be done.

Author Keywords
corpus linguistics, information visualization

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous

INTRODUCTION
Corpus linguistics is the study of language use by means of large electronic text collections, or corpora [2]. These are carefully compiled to ensure representativeness across desired features such as time, genre and the social status of writers/speakers. Nowadays corpora are often annotated for, e.g., part of speech or sentence structure; however, there is a lack of sophisticated tools for visualizing and analyzing such tagged and parsed data.

Information visualization is about using external tools to amplify cognition. Often these external tools are visual as more information is acquired through vision than via all the other senses combined [15, p. 2]. Visual and interactive representations of data improve problem solving and acquisition of insight.

Text as data is more challenging to visualize than numerical, nominal or categorical data. Text is high-dimensional and, e.g., equality tests are complicated because of multiple meanings and complex relations. Often the non-linguistically motivated text visualizations take shortcuts by ignoring the ordering relationships within the text, and by stemming the words (i.e., reducing them into their roots).

Here we will discuss our linguistically motivated visualizations for corpus linguistics. Although general-purpose visualization techniques provide a good starting point, techniques that dig deeper into the structure of the documents in the corpus, and work bottom-up from the texts, are needed to gain insight into linguistic variation and change.

Our work is based on the part-of-speech or POS-tagged version of the *Parsed Corpus of Early English Correspondence* (PCEEC) [9]. It is used as a running example in this paper whenever the copyright allows. The corpus consists of 4,968 letters written between the years 1415 and 1681, and has 2,155,446 words. The part-of-speech tagged text has each word marked up according to its definition and context, e.g., 'Hopkins_NPR' denotes that 'Hopkins' is a 'proper noun'. Although PCEEC is relatively small as a corpus, it is challenging to analyze because of the variations over time. In cases where the PCEEC copyright would be compromised, the freely available plain text version of *The Adventures of Sherlock Holmes* [3] by Sir Arthur Conan Doyle is used as example material.

TEXT VISUALIZATION
Text data comes in many forms: articles, books, novels, letters, web pages and blogs, just to name a few. In addition to texts created by a human author there are many computer-generated text genres as well, such as log files and other outputs from computer programs.

Text visualization is popular, both as an object of research and among the 'consumers of visualizations'. About one third of the user-created visualizations on the *Many Eyes* [8, 4] collaborative visualization service are related to text visualization, and media both in print and on the web routinely use such techniques as tag clouds and thematic maps to illustrate their texts (see, e.g., [12]).

The Many Eyes service has four text visualization modules in its selection (Figures 1 to 3 below). They provide a rep-

resentative sample of how the visualization community generates insight into text documents. Figure 1 shows a *Wordle* [5] visualization that illustrates the occurrence of words in PCEEC. In a Wordle visualization, the size of a word is determined simply by its frequency, and the placement of a word does not convey any additional information.

Figure 1. Wordle visualization of the two million words of the *Parsed Corpus of Early English Correspondence*.

The Wordle visualization in Figure 1 is impressive in that it manages to summarize almost 5,000 letters in just one picture. Although it is thought-provoking and entertaining, its worth from the linguistic point of view is questionable. As the designer of Wordle notes [14], a significant number of Wordle users do not even understand what the graphics are encoding, and a user might seek explanation for the proximity of certain words when they are just put together randomly. This is not to be taken as criticism of Wordle, as it does exactly what it was designed to do, and there are interesting applications for it. It has been used in the domain of text corpus visualization as well, to get an overview of Shakespeare's sonnets and plays. It might also be interesting to compare Wordle visualizations of PCEEC material per letter or per author, as in Feinberg's comparison of the inaugural addresses of the presidents of the United States [6]. In addition to Wordle, *Tag Cloud* is another word frequency visualization from the Many Eyes service. In a tag cloud the words of a text are laid out in alphabetical order and their size is in proportion to their frequency.

Figure 2 displays another text visualization, *Word Tree*, from the Many Eyes service which is a visual version of a traditional concordance (a list of words in their context). Instead of a traditional list view of such tools, the data is displayed as a tree. The primary difference besides the representation is that the tool displays only the right hand side of the current word's environment. However, selecting a word will permit exploration of the left hand side environment as well.

In a *Phrase Net* visualization, as seen in Figure 3, the user selects one of the pre-defined 'bigrams' or two-word patterns or creates a new one. The patterns define the elements that should appear between two words, and the tool creates a graph of all word pairs that match the pattern. Size is again used to encode frequency.

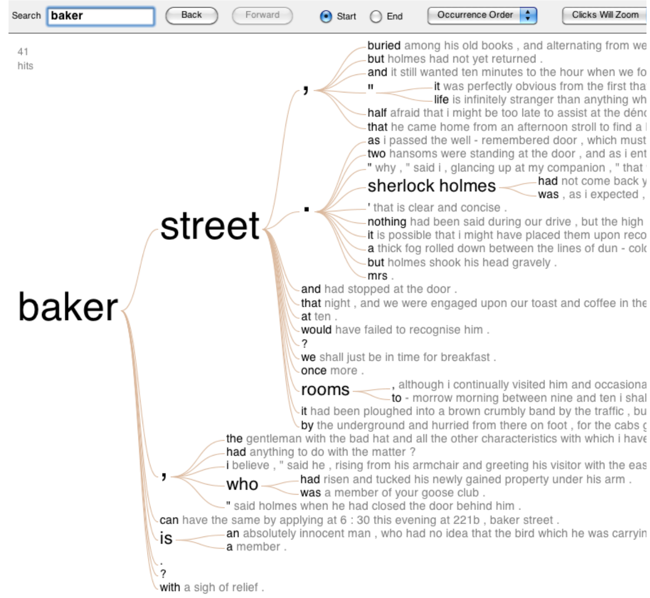


Figure 2. Word Tree visualization of *The Adventures of Sherlock Holmes* with focus on the word 'baker'.

SCENARIO

While Many Eyes can be a useful toolset for corpus linguistics, it is not designed to utilize annotation in corpora. This section presents a scenario of how the information visualization approach might be applied to solving a linguistically motivated problem in a POS-tagged corpus. Two open-source and freely available software tools are employed: the general statistical data-visualization system *Mondrian* [13] and the statistical system *R* [10].

Suppose that the aim of the study is to investigate if the 'nouniness' of language is affected by the sociolinguistic background variables as manifested in the PCEEC corpus. Nouniness, or the proportion of nouns in a text, can be determined in a part-of-speech tagged corpus by computing the percentage of words tagged as nouns against the whole corpus. The question of which of the tags are regarded as nouns is a matter of definition and open to some debate.

The first concern in a study of a POS-tagged text is data integrity, to make sure that the phenomenon under study is correctly encoded in the corpus. In this case, it means checking that the words tagged as nouns are really nouns, and that tokenization (how the text is segmented into words) is handled correctly. With a historical corpus, this step may involve many manual operations and computer scripts that 'prune' the corpus, and the result is a refined version of the corpus. Suppose the result from this step is in the following form (only 6 out of 2,154,210 lines are shown):

```
       word  tag
1       Mr.  NPR
2   Hopkins  NPR
3       yow  PRO
4 discourse  VBP
5    wisely  ADV
6       and  CONJ
```

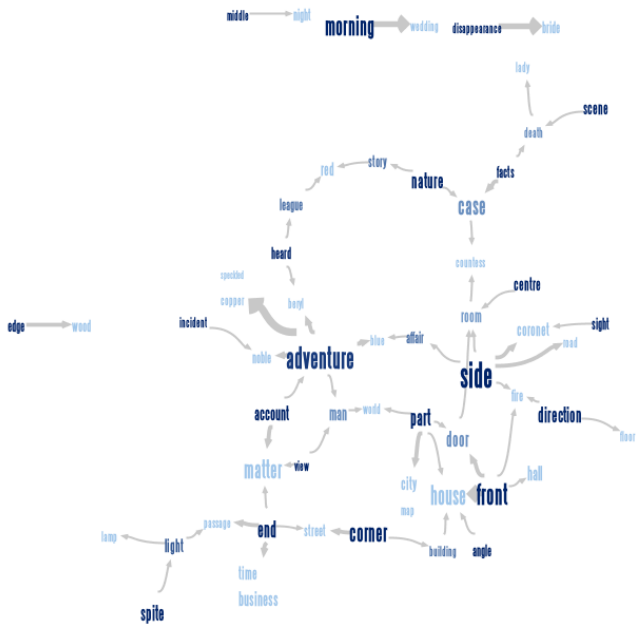


Figure 3. Phrase Net visualization of '* of the *' word patterns in *The Adventures of Sherlock Holmes*.

Once the word-category disambiguation is satisfactory, the corpus needs to be transformed into data tables for analysis. In this particular study, the text itself is not needed, only the frequencies of tags. The transformation can be done conveniently under R with the *reshape* package [16], which allows the user to 'melt' a data set into its components and then 'cast' it into a desired shape.

The desired statistical units in this study are letters, authors, years, time periods, and authors-by-time-periods. Summarizing the tag frequency data into these units can then be combined with the available sociolinguistic metadata, such as additional details about the authors and letters. Again, this can be carried out in R with its built-in relational operators and the aggregate function. After this step we will have the following tables:

```
pceecByText:  4968 obs.  of 101 variables
pceecByAuthor:  659 obs.  of 49 variables
pceecByYear:  243 obs.  of 26 variables
pceecByPeriod:  7 obs.  of 25 variables
pceecByAuthorPeriod:  729 obs.  of 50 variables
```

At this stage it is also trivial to extend the data tables with additional variables needed in the analysis, such as the word counts and the percentage of nouns.

As soon as the data tables exist, they can be explored with the Mondrian data visualization tool. Exporting the data tables in the current version of Mondrian is straightforward [13], and the forthcoming release can access R data frames directly from the R images.

A linguist might approach the data as follows. 'Nouniness' is a well-known indicator of formality in language use. Previous work in corpus-based sociolinguistics has shown that it varies depending on the gender of the speaker/writer: men tend to use more nouns than women [1, 11]. Furthermore, we know that people modify their language use depending on their interlocutors. Therefore, we might hypothesize that the linguistic formality of our material could vary depending on the gender of both the sender and the recipient of the letters. This hypothesis can be explored by looking at the histogram of the percentage of nouns in Mondrian, and then selecting the four possible combinations of the values of variables 'Sex.Sender' and 'Sex.Recipient' (Figure 4).

The visualization immediately provides strong support for our hypothesis. For instance, the letters sent by men to women (Sex.Sender='M' AND Sex.Recipient='F' – the highlight in the bottom right corner) have a clearly lower percentage of nouns than the letters sent by men to men (the top right corner). In fact, there seems to be a cline of formality: M2M $>$ M2F $>$ F2M $>$ F2F. This kind of analysis can be done rapidly in Mondrian-like interfaces by creating and combining persistent selections, or *selection sequences* in Mondrian terminology. The approach is based on a technique known as *brushing* [17] where the multiple, coordinated views propagate the selections of data items, thereby facilitating comparisons between the views.

In the mean time, back in the statistical system R, the findings from the explorative phase will undergo careful analysis to determine if a statistically significant difference exists, and the relevant graphics are re-created in print quality under R. The confirmed findings are then ready to be disseminated.

DISCUSSION

We have approached the information visualization needs and benefits of corpus linguists from two directions: from a sample of the 'state of the art' text visualizations, and from a scenario of how a linguistically motivated research problem might be tackled. Neither of these reflect realistically the current state of corpus-linguistic research as simple text concordancers and spreadsheet applications are still the prevailing tools. However, the use of the statistical system R is gaining popularity and is strongly endorsed by prominent computational linguists [7]. The downside of R is the steep learning curve and the dreaded command line interface.

What is crucial in visualization tools such as Mondrian from the linguist's standpoint is the chance for rapid and interactive exploration of data. It is known that interaction enhances discovery, and linguistic data visualizations are no exception. Generating "Aha! That's interesting!" exclamations may succeed with static data visualizations, such as Wordle tag clouds, but the chances are far better with interaction.

The problem we now have is how to make these tools more accessible to the linguistic community. With many extensive and versatile toolkits available, it makes sense to build on them instead of creating something from the ground up. In the case of R and Mondrian we are planning to improve the integration of the two systems and add new domain-specific data views as well.

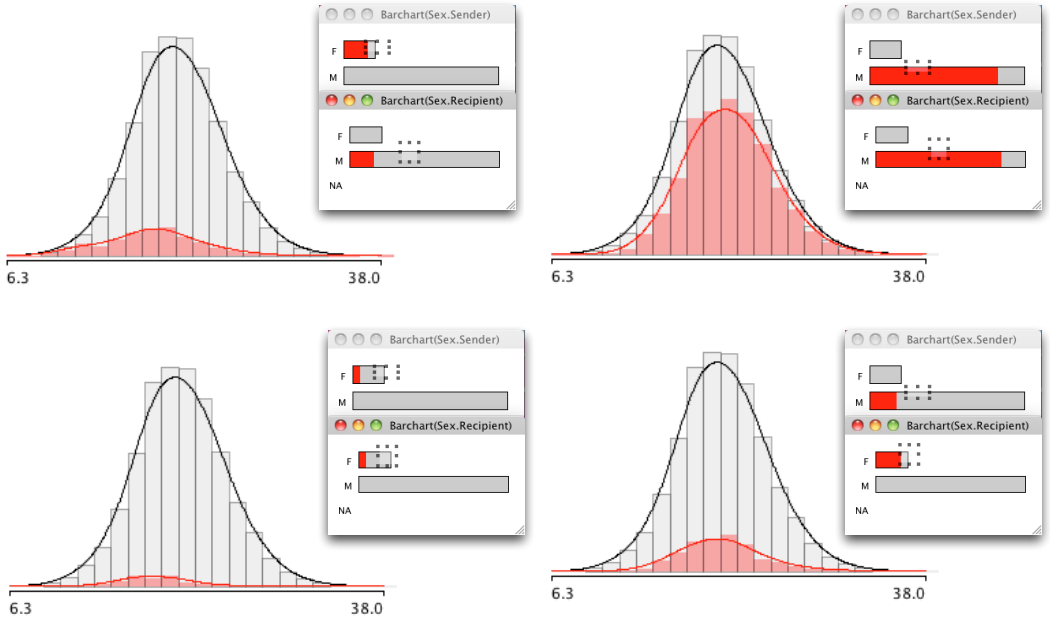


Figure 4. The effect of gender of letter sender/recipient on the 'nouniness' of text explored with the interactive data visualization system Mondrian.

REFERENCES

1. S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3):321–346, 2003.

2. D. Biber, S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

3. A. Conan Doyle, Sir. *The Adventures of Sherlock Holmes*, volume 1661 of *Project Gutenberg*. Project Gutenberg, 1999 [George Newnes Ltd, 1892].

4. C. M. Danis, F. B. Viegas, M. Wattenberg, and J. Kriss. Your place or mine? Visualization as a community component. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI Conference on Human factors in Computing Systems*, pages 275–284. ACM, 2008.

5. J. Feinberg, *Beautiful Word Clouds*, 2009. `http://www.wordle.net/`.

6. J. Feinberg, *Inaugural Addresses*, 2009. `http://www.research.ibm.com/visual/inaugurals/`.

7. S. T. Gries. *Quantitative Corpus Linguistics with R*. Routledge (Taylor and Francis), New York, 2009.

8. IBM Visual Communication Lab, *Many Eyes*, 2009. `http://manyeyes.alphaworks.ibm.com/manyeyes/`.

9. A. Nurmi, A. Taylor, A. Warner, S. Pintzuk, and T. Nevalainen. Parsed Corpus of Early English Correspondence (PCEEC), tagged version. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive, 2006.

10. R Development Core Team. R: A language and environment for statistical computing. `http://www.R-project.org/`, 2009.

11. P. Rayson, G. Leech, and M. Hodges. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1):133–152, 1997.

12. The New York Times Visualization Lab, 2009. `http://vizlab.nytimes.com/`.

13. M. Theus and S. Urbanek. *Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis)*. Chapman & Hall/CRC, 2008.

14. F. B. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1146, 2009.

15. C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufman, San Francisco, CA, second edition, 2004.

16. H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), November 2007.

17. G. Wills. Selection: 524,288 ways to say "this is interesting". In *InfoVis'96: Proceedings of the IEEE Symposium on Information Visualization 1996*, pages 54–61. IEEE Computer Society, 1996.

Visual Structured Summaries of Human Conversations

Giuseppe Carenini and Gabriel Murray
University of British Columbia
Vancouver, BC, Canada
carenini@cs.ubc.ca

ABSTRACT
This paper presents an interactive interface to create visually structured summaries of human conversations via ontology mapping. We have built highly accurate classifiers for mapping the sentences of a conversation in an ontology, which includes nodes for the Dialog Acts (DA) properties such as decision and subjective, along with nodes for the conversation participants. In contrast with previous work, our classifiers do not rely on features specific to any particular conversational modality. We are currently developing an interactive interface that allows the user to generate visual structured summaries by searching a conversation for sentences according to the ontology mapping. Our first prototype comprises two panels. The right panel displays the ontology, while the left panel of the our prototype displays the whole conversation, where sentences are temporally ordered. Given the information displayed in the two panels, the user can generate visual, structured summaries by selecting nodes in the ontology. As a result, the sentences that were mapped in the selected nodes will be highlighted. Our initial prototype builds on a component of the GATE system, which was originally developed as a tool for text annotation.

INTRODUCTION
Our lives are increasingly comprised of multimodal conversations with others. We email for business and personal purposes, attend meetings in person and remotely, chat online, and participate in blog or forum discussions. It is clear that automatic summarization can be of benefit in dealing with this overwhelming amount of interactional information. Automatic meeting abstracts would allow us to prepare for an upcoming meeting or review the decisions of a previous group. Email summaries would aid corporate memory and provide efficient indices into large mail folders.

The dominant approach to the challenge of automatic summarization has been *extraction*, where informative sentences in a document are identified and concatenated to form a condensed version of the original document. Extractive summarization has been popular at least in part because it is a binary classification task that lends itself well to machine learning techniques, and does not require a natural language generation component. There is evidence that human abstractors at times use sentences from the source documents nearly verbatim in their own summaries, justifying this approach to some extent [9]. Extrinsic evaluations have also shown that, while extractive summaries may be less coherent than human abstracts, users still find them to be valuable tools for browsing documents [7, 10, 13].

However, these same evaluations also indicate that concise abstracts are generally preferred by users and lead to higher objective task scores. The limitation of a cut-and-paste summary is that the end-user does not know *why* the selected sentences are important; this can often only be discerned by exploring the context in which each sentence originally appeared. One possible improvement is to create *structured extract summaries* that represent an increased level of abstraction, where selected sentences are grouped according to the entities they mention as well as to phenomena such as *decisions*, *action items* and *subjectivity*, thereby giving the user more information on why the sentences are being highlighted. For example, the sentence *Let's go with a simple chip* is about a *simple chip* and represents both a decision and the expression of a positive subjective statement.

While much attention in recent years has been paid to (unstructured) extractive summarization of human conversations, including meetings [5], emails [17, 2], telephone conversations [21] and internet relay chats [20], in this paper we present a novel approach to generating visual, structured summaries of human conversations. In our approach sentences are first mapped to nodes in a conversation ontology. Then, the user can search the conversation through an interactive visualization that effectively display both the ontology and the conversation, and allows the user to search the conversation based on the ontology mapping.

The mapping of sentences to the ontology is performed by first identifying all the entities referred to in the conversation, and then by utilizing classifiers relating to a variety of sentence-level phenomena such as *decisions*, *action items* and *subjective sentences*. We achieve high classification accuracy by using a very large feature set integrating conversation structure, lexical patterns, part-of-speech (POS) tags and character n-grams.

Once the mapping is created the user can generate visual, structured summaries by searching a conversation for sen-

tences that convey information about nodes in the ontology. These sentences are highlighted in the context of the whole conversation. For instance, if a user wanted to highlight all the sentences in an email thread expressing *decisions* on the *remote control* made by the *project manager*, she could achieve that by simply selecting the corresponding nodes in the ontology.

In this paper, we first describe the process of mapping sentences to a conversation ontology and then we present our interface to generate visual structured summaries.

ONTOLOGY MAPPING

Our approach relies on a simple conversation ontology. The ontology is written in OWL/RDF and contains two core upper-level classes: Participant and Entity. When additional information is available about participant roles in a given domain, Participant subclasses such as ProjectManager can be utilized. The ontology also contains six properties that express relations between the participants and the entities. For example, the following snippet of the ontology indicates that *hasActionItem* is a relationship between a meeting participant (the property domain) and a discussed entity (the property range).

```
<owl:ObjectProperty rdf:ID="hasActionItem">
  <rdfs:domain rdf:resource="#Participant"/>
  <rdfs:range rdf:resource="#Entity"/>
</owl:ObjectProperty>
```

Similar properties exist for decisions, actions, problems, positive subjective sentences, negative subjective sentences and general extractive sentences (important sentences that may not match the other categories), all connecting conversation participants and entities. The goal is to populate the ontology with participant and entity instances from a given conversation and determine their relationships. This involves identifying the important entities and classifying the sentences in which they occur as being decision sentences, action item sentences, etc.

Our current definition of entity is simple. The entities in a conversation are noun phrases with mid-range document frequency. This is similar to the definition of concept as defined by Xie et al. [19], where n-grams are weighted by *tf.idf* scores, except that we use noun phrases rather than any n-grams. We use mid-range document frequency instead of *idf* [4], where the entities occur in between 10% and 90% of the documents in the collection. We do not currently attempt coreference resolution for entities; recent work has investigated coreference resolution for multi-party dialogues [11, 6], but the challenge of resolution on such noisy data is highlighted by low accuracy (e.g. F-measure of 21.21) compared with using well-formed text (e.g. monologues).

We map sentences to our ontology's object properties by building numerous supervised classifiers trained on labeled decision sentences, action sentences, etc. A general extractive classifier is also trained on sentences simply labeled as important. After predicting these sentence-level properties, we consider a participant to be linked to an entity if the participant mentioned the entity in a sentence in which one of these properties is predicted. We give a specific example of the ontology mapping using this excerpt from the AMI corpus [3]:

1. A: And you two are going to work together on a *prototype* using *modelling clay*.
2. A: You'll get *specific instructions* from your *personal coach*.
3. C: Cool.
4. A: Um did we decide on a *chip*?
5. A: Let's go with a *simple chip*.

Example entities are italicized. Sentences 1 and 2 are classified as action items. Sentence 3 is classified as positive-subjective, but because it contains no entities, no $< participant, relation, entity >$ triple can be added to the ontology. Sentence 4 is classified as a decision sentence, and Sentence 5 is both a decision sentence and a positive-subjective sentence (because the participant is advocating a particular position). The ontology is populated by adding all of the sentence entities as instances of the Entity class, all of the participants as instances of the Participant class, and adding $< participant, relation, entity >$ triples for Sentences 1, 2, 4 and 5. For example, Sentence 5 results in the following two triples being added to the ontology:

```
<ProjectManager rdf:ID="participant-A">
<hasDecision rdf:resource="#simple-chip"/>
</ProjectManager>

<ProjectManager rdf:ID="participant-A">
<hasPos rdf:resource="#simple-chip"/>
</ProjectManager>
```

We have tested our classifiers both on meeting and email data, the AMI [3] and BC3 [18] corpus respectively. On meetings, we achieve remarkable performances, with classification AUROCs ranging from .93 to .77, depending on the classification task. On emails, results are slightly lower, but still potentially useful, with classification AUROCs ranging from .75 to .66. For a detailed discussion of the results see [12] [1].

A key feature of our mapping approach is that it only relies on generic conversational features and can therefore be applied to a multi-modal conversation, for instance a conversation that spans both an email thread and a meeting. Noticeably, our classifiers achieve similar results to [8], [15, 14], [16], who perform these classification tasks by relying on meeting-specific or email-specific features (e.g., prosody for meetings).

GENERATING VISUAL STRUCTURED SUMMARIES

We are developing an interactive interface that allows the user to generate visual structured summaries by searching a conversation for sentences according to the ontology mapping. Figure 1 shows our first prototype for such an interface. The right panel displays the ontology which includes, at the time of writing, nodes for the Dialog Acts (DA) properties such as decision and subjective, along with nodes for

[1] If this paper is not be accepted to NAACL, a draft version can be requested to the authors.

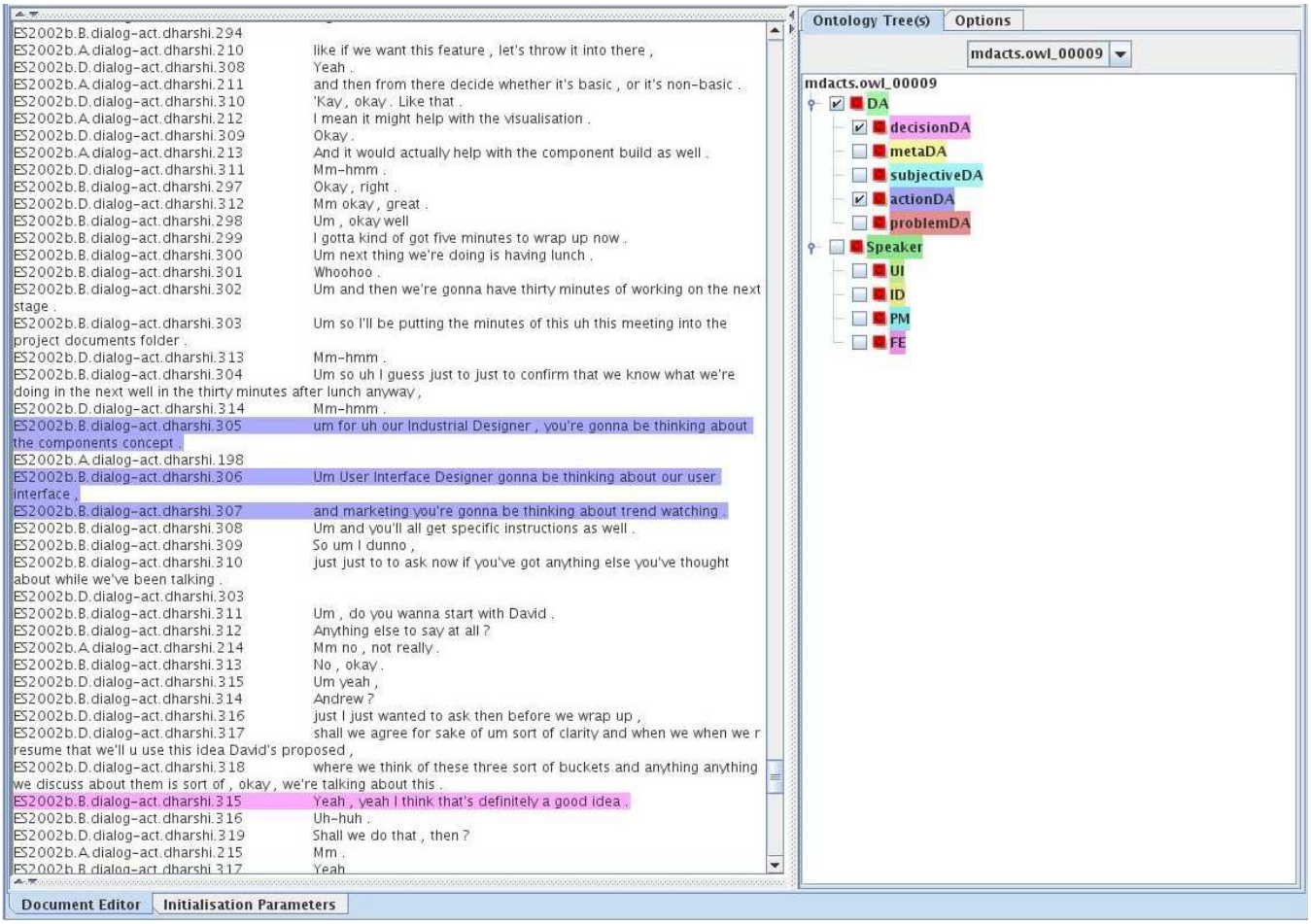


Figure 1. Screenshot of our interface for creating visual, structured summaries of human conversation

the conversation participants (Speaker in the figure) [2]. The left panel of the interface displays the whole conversation, where sentences are temporally ordered. Given the information displayed in the two panels, the user can generate visual, structured summaries by selecting nodes in the ontology. As a result, the sentences that were mapped in the selected nodes will be highlighted.

For instance, the left panel in Figure 1 displays a sample meeting from the AMI corpus whose sentences have been classified and mapped in the conversation ontology. In the example, since the user has selected the nodes *decision* and *action* in the ontology, the sentences mapped in those nodes are highlighted in the context of the whole conversation. In the current interface each node is associated with a different color and a sentence mapped into multiple selected nodes is colored as the "intersection" of the corresponding colors. This solution is not satisfactory and we are investigating more effective techniques to visually convey this information.

[2]We are currently adding to the interface nodes for all the entities extracted from the conversation (as described in the previous section).

Our initial prototype builds on a component of the GATE system [1], which was originally developed as a tool for text annotation.

CONCLUSIONS AND FUTURE WORK

This paper presents an interactive interface to create visually structured summaries of human conversations via ontology mapping. So far, we have built highly accurate classifiers for the mapping phase, that, in contrast with previous work, do not rely on features specific of any particular conversational modality. We have also implemented a first prototype of the interface that display both the ontology and the conversation, and allows the user to search the conversation based on the ontology mapping.

In the near future we plan to complete the development of the prototype. First, we are currently extending the displayed ontology to also include the entities mentioned in the conversation. Second, we will study how to effectively highlight sentences that were mapped to multiple nodes in the ontology. Once the summarization interface is completed, we intend to perform an extrinsic evaluation, in a way similar to [7, 10, 13].

REFERENCES

1. K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10, 2004.

2. G. Carenini, R. Ng, and X. Zhou. Summarizing email conversations with clue words. In *Proc. of ACM WWW 07, Banff, Canada*, 2007.

3. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39, 2005.

4. K. Church and W. Gale. Inverse document frequency IDF: A measure of deviation from poisson. In *Proc. of the Third Workshop on Very Large Corpora*, pages 121–130, 1995.

5. M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372, 2006.

6. S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky. Resolving *"You"* in multi-party dialog. In *Proc. of SIGdial 2007, Antwerp, Belgium*, 2007.

7. L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *Proc. of ACM MULTIMEDIA '99, Orlando, FL, USA*, pages 489–498, 1999.

8. P.-Y. Hsueh, J. Kilgour, J. Carletta, J. Moore, and S. Renals. Automatic decision detection in meeting speech. In *Proc. of MLMI 2007, Brno, Czech Republic*, 2007.

9. J. Kupiec, J. Pederson, and F. Chen. A trainable document summarizer. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA*, pages 68–73, 1995.

10. K. McKeown, J. Hirschberg, M. Galley, and S. Maskey. From text to speech summarization. In *Proc. of ICASSP 2005, Philadelphia, USA*, pages 997–1000, 2005.

11. C. Muller. Resolving *It*, *This* and *That* in unrestricted multi-party dialog. In *Proc. of ACL 2007, Prague, Czech Republic*, 2007.

12. G. Murray and G. Carenini. Interpretation and transformation for abstracting conversations. In *Submitted to the 2010 North American ACL*, 2010.

13. G. Murray, T. Kleinbauer, P. Poller, S. Renals, T. Becker, and J. Kilgour. Extrinsic summarization evaluation: A decision audit task. In *Proc. of MLMI 2008, Utrecht, the Netherlands*, 2008.

14. G. Murray and S. Renals. Detecting action items in meetings. In *Proc. of MLMI 2008, Utrecht, the Netherlands*, 2008.

15. M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. Detecting and summarizing action items in multi-party dialogue. In *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium*, 2007.

16. S. Raaijmakers, K. Truong, and T. Wilson. Multimodal subjectivity analysis of multiparty conversation. In *Proc. of EMNLP 2008, Honolulu, HI, USA*, 2008.

17. O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen. Summarizing email threads. In *Proc. of HLT-NAACL 2004, Boston, USA*, 2004.

18. J. Ulrich, G. Murray, and G. Carenini. A publicly available annotated corpus for supervised email summarization. In *Proc. of AAAI EMAIL-2008 Workshop, Chicago, USA*, 2008.

19. S. Xie, B. Favre, D. Hakkani-Tür, and Y. Liu. Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization. In *Proc. of Interspeech 2009, Brighton, England*, 2009.

20. L. Zhou and E. Hovy. Digesting virtual "geek" culture: The summarization of technical internet relay chats. In *Proc. of ACL 2005, Ann Arbor, MI, USA*, 2005.

21. X. Zhu and G. Penn. Summarization of spontaneous conversations. In *Proc. of Interspeech 2006, Pittsburgh, USA*, pages 1531–1534, 2006.

Visual Abstraction and Ordering in Faceted Browsing of Text Collections

VinhTuan Thai, Siegfried Handschuh
Digital Enterprise Research Institute, National University of Ireland, Galway
firstname.lastname@deri.org

ABSTRACT
While faceted navigation interfaces can assist users in exploring an information collection, there is yet little support for users in choosing a relevant item from the set of items returned from a filtering process. In this paper, we propose using a multi-dimensional visualization as an alternative to the linear listing of focus items. We describe how visual abstraction based on a combination of structural equivalence and conceptual structure can be used to deal with a large number of items, as well as visual ordering based on the importance of facet values to support cross-facets comparison of focus items. This visual support for faceted browsing has been developed for visual exploration of text collections.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Graphical user interfaces

FACETED NAVIGATION - PROS AND CONS
Faceted navigation is a proven technique for exploration and discovery of an information collection [3]. Apart from being well-studied in many research work e.g. [12, 4, 2], its usefulness is attested by the popular uptake in many commercial websites. At the core of faceted navigation is a set of flat or hierarchical facets, which are categories characterizing items in a large collection [3]. Each facet has one or more facet values and each item may be associated with a subset of these values [1]. As such, this navigation paradigm requires rich metadata expressing relationships between facet values and information items. Users' selections of facet values result in either conjunctive or, more commonly, disjunctive queries executed on the dataset. The matching items (or focus items [1]) are then displayed as results of the filtering process.

We were interested in user experiences with existing faceted user interfaces (UIs), therefore we invited three persons who were familiar with faceted browsing to participate in **contextual interviews**. They were asked to demonstrate their recent use of a faceted UI to achieve a real task of their own. One accessed the www.komplett.ie website to buy a PCI-Express graphic card, one used the www.yelp.com website to look for Vietnamese restaurants in New York, and the last one went to www.daft.ie to find a room in Galway whose rent should cost less than 400 euros. They explained their interactions while we were observing. Afterwards, each of them discussed what were good about these websites and what could be improved. The users' feedback is as follows:

- The facets in these websites were highly appreciated as they were relevant and helpful to narrow down the search. One subject in particular was not happy with the default set of facets, as they did not reflect his most relevant criteria, therefore he always used the "Advanced Search" option which provided all available facets.
- Comparison of items across different facets is very important for making decisions. To avoid missing the best matches, users had to look at different combinations of facet values inherent in the focus items. Sorting by one facet at a time was thus not considered effective.
- Facets are not equally important. Some facets are more important than others (e.g. neighborhood can be more important than room type).
- Having to go through a long list of focus items is time-consuming. While disjunctive queries allow displaying a wider set of items matching one or more values of a facet, users needed to look into the details of each item to figure out which values of a facet an item matched. This is in line with results from a study on faceted UI [5], which showed that while users spent equally much time on the query, the facets and the results on the first results page, they focused entirely on the items on the second and third results pages. This suggested that after choosing the facets to narrow down their search, users still needed to spend a considerable amount of time to look for a specific item.

The above feedback begs the question if properties of focus items can be visually displayed in such a way that users can make a better informed decision faster than having to traverse all pages of results and looking at one item after another. This question is also relevant to visual text analysis research. While the websites chosen above by the subjects were commercial websites catering for different products, they shared many similarities with faceted UIs for text collections, such as the one studied in [5]. This means that in faceted browsing of texts, filtering interactions also result in document items returned in a list, usually sorted by overall relevance, and users still have to traverse through many results pages to select a particular document for further anal-

Workshop on Intelligent Visual Interfaces for Text Analysis (IVITA2010), ACM IUI 2010, February 7, 2010, Hong Kong, China.

ysis. The key limitation is the lack of an aggregated view showing how each focus item relates to each of the facet values. While certain issues were raised about faceted UIs [3], they tend to focus on the display of a large number of facets, e.g. which facets to show (adaptively) when there are many. The issues identified here regarding focus items representation have largely been left untouched. Based on the users' feedback, we argue that their experiences with faceted UIs can be improved if the following design desiderata are met:

- Each focus item should have a compact representation expressing its correspondences to facet values.
- Users should be able to perform cross-comparison of focus items over different values of one or more facets.
- The display can be visually abstracted to deal with a large amount of focus items.
- Users should be able to interactively reorder facets based on their preferences, resulting in different displays of focus items.

As such, we propose using a multi-dimensional visualization, one dimension for each facet value, as an alternative to the linear result listing paradigm. In the rest of the paper, we provide the context and then the proposed solution.

CONTEXT

Our proposed solution is developed within the context of a filtering mechanism to support exploration of a text collection in a personalized manner. The design of the initial prototype IVEA [10] is based on Shneiderman's visual information-seeking mantra "*Overview first, zoom and filter, then details-on-demand*" [8]. It employs multiple coordinated views to guide users through this workflow. The core element in IVEA is a simple user-defined ontology encapsulating their sphere of interest. Various statistics about the relationships between documents and entities in the ontology are visually presented to facilitate the exploration.

PROPOSED VISUALIZATION

We proposed in [11] the initial idea on a matrix-based multi-dimensional visualization, which was inspired by FOCUS [9] and TableLens [7], to show the correspondences between documents and a taxonomy of entities of interest to users. This visualization, in effect, supports filtering similarly to faceted navigation in that users can select entities from hierarchical facets and then documents relevant to those entities are displayed. Since hierarchical relationships between entities are taken into account, selecting a class will result in the automatic inclusion of all of its direct instances and recursively, all of its subclasses. Thus, facet selection for filtering can be done at different levels of granularity and multiple facet values can be selected in a single operation. In the matrix, rows represent selected entities, columns represent documents containing at least one of those entities, and each cell shows the relevance value (TF-IDF based score) of a document with respect to an entity via its height. Here each entity is linked to a user-defined set of associating terms. As such, abbreviations, linguistic variations, conceptually related terms and synonyms can be taken into account. This is important, since in many cases, documents' authors adopt a rhetorical writing style by choosing different wordings, and use them as a semantic camouflage with the intended purpose of influencing readers into accepting misleading interpretations of the information being presented.

To decide a cell's height, we use k-means clustering to identify three clusters of relevance values, and the maximal values of the three clusters are used as thresholds. The vertical part of the cross-hair highlighter helps to focus on which entities a document contains and its horizontal part helps to show the distribution of an entity in a collection. Users can also remove facet values by right-clicking on an entity and the whole respective row is removed from the visualization.

Although it meets two of the design desiderata, the visualization provides no visual abstraction to cater for a large number of documents and no interactive ordering of facet values to enable users to easily compare items across different facet values (entities). Next we present the proposed solution. Here, documents are information items and concepts/entities of interest to users are facet values.

VISUAL ABSTRACTION OF DOCUMENTS

Relationships between documents and a set of concepts can be represented by a bipartite graph $G = (D, C, E)$ whose vertices belong to two disjoint sets D representing documents and C representing concepts. If $d_i \in D$ is relevant to $c_j \in C$, then there is an edge $(d_i, c_j) \in E$ connecting them, whose weight is the relevance of d_i with respect to c_j.

Given the typically available screen resolution, too much data is confusing and limiting information manipulation. Therefore, it is important to collapse visual information when desired so that it takes up less screen space and users can focus on what is being shown more effectively. This need is equivalent to interactively collapsing and expanding vertices in the set of documents and the set of concepts so that the view is collapsed and expanded accordingly. In this respect, two mechanisms are provided: semantic zooming and document grouping, which can be combined in use.

The semantic zooming feature provides different levels of abstraction based on the hierarchical relationships among entities (facet values). The hierarchy attached with the matrix allows users to dynamically drill down or roll up to achieve views at different conceptual levels of detail, as indicated by the collapsible glyph in Fig.1, to focus on particular subgroups. This, in effect, is the aggregation of vertices in the set of concepts C, hence replaces edges that connect documents to instances of a class with edges that connect documents to that class only. For instance, the bigraph in Fig.2 shows the relationships between 5 documents and 4 concepts. Assuming that concepts c_1, c_2, c_3 are instances of a class, they therefore can be grouped into a single concept c_{123} representing that class. Any document that contains any or all of the three concepts c_1, c_2, c_3 is considered as containing the concept c_{123}. Thus, the resulting bigraph has fewer edges, as shown in Fig.2, whereby all documents now connect to the concept c_{123} instead of to the three concepts individually.

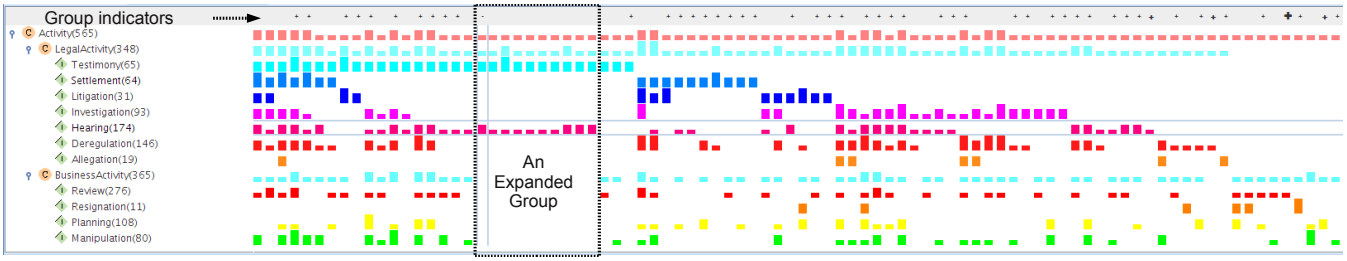


Figure 1. Document Grouping

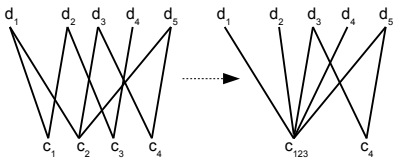


Figure 2. Semantic Zooming Bigraphs

While semantic zooming can abstract away a lot of details, the number of documents in a relatively large collection can still be too much to be effectively displayed on a limited screen space. Here the document grouping feature provides further abstraction based on the notion of Structural Equivalence of individuals in social networks [6], defined as below.

Definition Objects a, b of a category C are *structurally equivalent* if a relates to every object x of C in exactly the same way as b does [6].

This notion is used to partition objects in a set into classes of structurally equivalent objects, which leads to the ability to derive a reduced set of categories in which belonging objects are considered equivalent. When these objects are individuals in a social network, the set of derived categories represents "maximal relationally homogeneous groups" [6]. Here we treat documents and concepts as objects and adapt the Structural Equivalence notion as per below.

Definition Given a set of concepts $C = \{c_1, .., c_n\}$, the set of structurally equivalent documents with respect to C consists of documents that contain all elements of C.

In other words, documents d_i and d_j in D are structurally equivalent with respect to C if there exist edges (d_i, c_k) and $(d_j, c_k) \in E$, for $k = (1, .., n)$. As such, given a set of entities C , we can identify a set of structurally equivalent documents with respect to this set and treat them as a group. This, however, has two limitations: (1) The requirement for documents to be structurally equivalent is strict in that they need to contain **all** elements of a given concept set, therefore the number of documents satisfying this requirement will not likely be large and (2) only one group can be identified given a set of entities, which is not significantly helpful in dealing with a large collection. Thus, our approach is to consider the powerset of C (excluding the empty set) and find groups of structurally equivalent documents with respect to each of those subsets. For example, there are four concepts c_1, c_2, c_3, c_4 , and five documents relevant to them in such a way that is represented by the left bigraph in Fig.3. Here documents d_1, d_3, d_5 contain c_2, c_4 and not any other concepts. Therefore, these three documents can be put together into a group d_{135} as shown in the right bigraph.

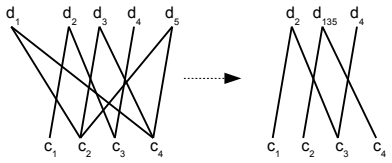


Figure 3. Document Grouping Bigraphs

This approach enables more flexibility with regard to the levels of granularity at which information is viewed and manipulated. As in Fig.1, although the screen space is limited, the visualization can still cope with a large set of documents. Here, the filtering process results in 565 relevant documents. However, since 50 structurally equivalent groups are identified, only 76 columns need to be shown, as only one (randomly chosen) document of a group is initially displayed on the matrix, while other documents that do not belong to any groups are still displayed as a regular column each. In fact, if there are n selected facet values, only a maximum of $2^n - 1$ columns are needed for the initial display. The column of the representative document of a group has a '+' sign on top, which is a visual cue to indicate that there are more documents containing exactly the same set of entities. We also use k-means clustering on different group sizes to find three clusters of sizes and use the maximal values of the three clusters as thresholds. Thus, the size of the '+' sign can indicate the relative size of a group. Hovering over a '+' sign will pop-up the exact number of documents in that group. This visual cue overcomes the need to show numeric values, which require varying spaces and can distort the consistent layout of the matrix. Clicking on this representative column will make visible all documents in a group and its visual cue changes to '-' as shown in Fig.1. Clicking again on the representative column will hide other documents in that group. This "focus+context" interaction simplifies the comprehension of the visual display of a large number of documents without users having to examine a matrix containing a large number of columns, since the initial display does not depend on the actual number of focus items (documents).

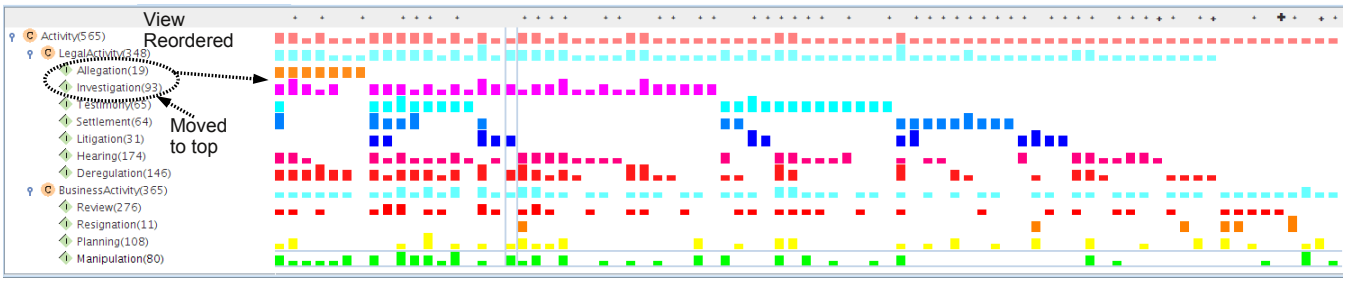


Figure 4. Facet Reordering

VISUAL ORDERING BASED ON FACET VALUES

As previously mentioned, not all facets are equally important. For each user, there is an order of importance of facet values accordingly. Therefore, we consider facet values that are placed on top are more important than those below them. Thus, documents and groups of documents are reordered based on their correspondences with facet values. As in Fig.1, in the facet "LegalActivity", the value "Testimony" has the highest position, therefore documents and groups of documents that are relevant to "Testimony" are moved to the left and those that do not are moved to the right. Within these two groups, they are subsequently ordered by their relevances to the second value, "Settlement". This ordering is done similarly until the last facet value. This ordering can be efficiently achieved using a bitmap of a document or a group of documents, which is constructed by assigning a 1 bit if a document/group of documents is relevant to a facet value, a 0 bit otherwise, and the first facet value corresponds to the highest order bit. For instance, the left-most document in **Fig.1** corresponds to the value "11111111011001", highest among the derived bitmap values. Furthermore, users can interactively reorder facet values while exploring a text collection. As shown in Fig.4, in the facet "LegalActivity", if the values "Allegation" and "Investigation" are considered more important, they can be moved (via drag-and-drop) on top. The view is changed accordingly as a result. We believe that this visual ordering based on facet values enables users to easily compare focus items, in this case documents/groups of documents, across facet values in a meaningful way.

DISCUSSION

Initial users' feedback on this visualization support for faceted browsing has been positive. They do acknowledge the need to get used to it, due to its differences from the more familiar display paradigm of linear listing of focus items. Although this is developed in the context of documents as information items, we believe it can be applied to other kinds of information items, since apart from temporal and spatial facet values which require separate timeline and map displays, most facet values are categories that can be visually encoded using iconic representations as in the proposed visualization.

ACKNOWLEDGMENTS

This work is funded in part by Science Foundation Ireland under Grant No.SFI/08/CE/I1380 (Lion-2) and the EU under Grant No.FP7-ICT-2007-1-216048 (FAST project).

REFERENCES

1. E. C. Clarkson, S. B. Navathe, and J. D. Foley. Generalized formal models for faceted user interfaces. In *Proc JCDL '09*, pages 125–134, 2009.
2. M. Dörk, S. Carpendale, C. Collins, and C. Williamson. Visgets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Trans. on Visualization and Computer Graphics*, 14:1205–1212, 2008.
3. M. A. Hearst. UIs for Faceted Navigation: Recent Advances and Remaining Open Problems. In *HCIR'08*, 2008.
4. D. Huynh, D. Karger, and R. Miller. Exhibit: Lightweight Structured Data Publishing. In *Proc WWW'07*, 2007.
5. B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *Proc JCDL '09*, pages 313–322, 2009.
6. F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1:49–80, 1971.
7. R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proc CHI'94*, pages 318–322, 1994.
8. B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Visual Languages*, pages 336–343, 1996.
9. M. Spenke, C. Beilken, and T. Berlage. FOCUS: the interactive table for product comparison and selection. In *Proc UIST '96*, pages 41–50, 1996.
10. V. Thai, S. Handschuh, and S. Decker. IVEA: An Information Visualization Tool for Personalized Exploratory Document Collection Analysis. In *Proc ESWC 2008*, pages 139–153, 2008.
11. V. Thai, S. Handschuh, and S. Decker. Tight coupling of personal interests with multi-dimensional visualization for exploration and analysis of text collections. In *Proc IV08*, pages 221–226, 2008.
12. K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proc CHI'03*, 2003.